공학석사학위논문

글에 담긴 정서 인식을 위한 분자 컴퓨터 모델

Molecular computational modeling of recognizing emotion in text

2005년 8월

서울대학교 대학원

협동과정 인지과학전공

손 정 욱

글에 담긴 정서 인식을 위한 분자 컴퓨터 모델

Molecular computational modeling of recognizing emotion in text

지도교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2005년 4월

서울대학교 대학원

협동과정 인지과학 전공

손 정 욱

손정욱의 공학석사 학위논문을 인준함

2005년 6월

심사위원장 고 성 룡 (印)

부 위 원 장 장 병 탁 (印)

심 사 위 원 신 효 필 (印)

초 록

공동체에서 구성원들 사이의 정서적 커뮤니케이션은 매우 중요하다. 컴퓨터 같은 지능적인 기계들은 이미 우리 생활에서 필수적인 역할을 담당하게 된 지 오래지만 이들의 정서적 의사소통 능력은 거의 없는 것이나 마찬가지이다. 본 연구는 이러한 맥락에서 글에서 정서적 정보를 인식해내는 모델을 제시한다.

단어의 의미적인 관련성과 정서의 인식은 서로 밀접한 관련이 있다는 심리학적 연구를 바탕으로 본 모델은 의미 네트워크를 구성하고 이를 통해 텍스트의 정서를 추론해낸다. 사람의 인지과정을 모사하는데 분자 컴퓨터를 이용한 모델링은 초병렬적인 계산 특성과 함께 생물학적 신호체계와 유사하다는 장점을 가진다. 분자 컴퓨터를 이용한 계산 과정은 DNA 분자들이 연쇄적으로 자기조립되어 결과 구조물을 생성해내는 과정으로 모델의 최종 출력은 PLM 알고리즘에 의해 분자들의 통계적 특성을 통해 계산된다. 단어를 나타내는 DNA 분자와 각각의 단어들을 연결시켜주는 DNA 분자들 사이의 미시적인 자기조립 과정은 단순한 논리적 추론에 가깝지만, 거시적인 모델의 처리 과정은 의미 네트워크를 통한 확률적인 계산 과정이다.

텍스트는 전처리 과정을 거쳐서 핵심어를 추출하여 단어 주머니 형태로 변환되어 모델의 입력으로 사용되는데 의미적으로 관련된 단어들을 의미 네

트워크를 이용해 걸러내고 이를 이용해 적절한 정서를 추론하게 된다. 본 연구에서 사용된 네트워크의 크기는 5000 단어이고 약 30만개의 문장을 통해 학습되었다. 모델이 산출한 정서는 사람의 평가에 의해 정확도를 측정하였고 텍스트의 종류에 따라 약 65%에서 80%의 정확도를 보여주었다.

주제어: 분자컴퓨터, DNA 컴퓨터, PLM, 정서 인식, 의미 네트워크

학번: 2003-23195

목 차

그 림 목 차

1. 서 론

사람이 공동체의 구성원으로 원만하게 살아가기 위해서 가장 중요한 것은 정서적 능력이다. 다른 구성원들의 감정을 올바르게 인식하고, 이해하고, 구성원들에게 적절하게 표현하는 것이 바로 그것인데, 전통적인 수학적 혹은 언어적 지능 보다 이러한 정서지능(emotional intelligence)이 성공적인 삶에서 더 중요한 역할을 한다는 것은 이제는 널리 알려진 사실이다.

의사 소통 과정에 정서적 정보를 담거나 알아채는 사람의 능력은 사람과 사람 사이의 의사 소통 과정에서뿐만 아니라 사람과 기계 사이의 의사 소통 과정에서도 자연스럽게 일어난다. 가령 컴퓨터 프로그램이 원하는 대로 동작하지 않으면 사람들은 당황스러움이나 짜증을 느낄 것이다. 게다가 오랜 시간을 걸려 작업한 결과물이 일순간에 망가져 버린다면 심한 분노를 느낄지도 모른다. 상대가 사람의 기분은 알 수 없는 단순한 기계 덩어리에 불과하다는 것을 알고 있음에도 그러한 감정은 사람에 대해서 표출되는 것처럼 기계에 대해서도 똑같이 표출된다.[1]

과학 기술의 눈부신 발달로 사람들은 컴퓨터와 로봇 같은 지능적인 기계들에 둘러싸여 지내는 시간이 많아졌다. 일상 생활에서도 이러한 지능적인 기계들이 필수적인 역할을 담당하게 된 지 오래지만 불행히도 이들의 정

[1] [Picard, R. W. (1997)]

서적 의사 소통 능력은 거의 없는 것이나 마찬가지이다. 단순한 계산 성능에서는 이미 컴퓨터가 사람의 그것을 앞지른 부분이 있음에도 대부분의 사람들에게 기계는 차가운 기계일 뿐이다.

최근의 컴퓨터 프로그램들은 사용자에게 좀더 친근한 환경을 제공해주기 위해 많은 노력을 기울이고 있다. 과거의 문자 기반의 딱딱한 인터페이스에 비하면 현재의 프로그램 인터페이스는 놀라운 발전이다. 하지만 대부분의 이런 노력들은 아바타(avatar)나 대화식(interactive) 도우미 등을 이용하여 프로그램의 표현력을 높이는 데에 주로 초점을 맞춰 왔다. 정서적 의사소통 과정이 정서 정보의 흐름이 양방향으로 균형적으로 이루어질 때 가장 큰 효과를 가진다는 점에서 볼 때, 사람에서 컴퓨터로의 정서 정보의 흐름이 상대적으로 부족한 것이다. 컴퓨터에 정서지능(emotional intelligence)을 부여하려는 이러한 시도에서 정서의 인식이 올바르게 이루어져야 적절한 정서적 반응을 보여줄 수 있기 때문에 정서의 인식은 정서의 표현보다 선행되어야 하는 중요한 부분이다.

사람의 두뇌에는 전기적 신호 전달 과정과 더불어 화학적 신호 전달 과정도 존재한다. 신경전달물질(neurotransmitter)과 신경조절물질(neuroregulator)들이 관여하는 이러한 화학적 처리과정이 단적으로 드러나는 예는 사람의 감정이다.[2] 많은 연구에서 특정 호르몬이나 신경전달물질이 특정 감정에 밀접한 관련이 있음을 밝혀냈고, 심리학과 신경생리학의 최근 연구 결과들은

[2] 정서와 neurochemistry에 관한 내용은 [Panksepp, J. (1998)] 참조.

정서는 인지과정과 결코 따로 떼어서 생각할 수 없는 것임을 보여준다.[3] 따라서 정서를 비롯한 사람의 인지과정을 이해하기 위해서는 뉴런의 전기적 신호 전달에 대한 기존의 연구뿐만 아니라 신경화학적(neurochemistry) 관점의 새로운 접근 방법이 필수적이다.

본 연구에서는 정서 인식 모델에 화학적 메커니즘에 기반하는 분자컴퓨터(molecular computer)를 이용한다. 뉴런의 전기적 신호전달을 모사한 기존의 신경망 모델에 비해 분자컴퓨터는 초병렬적인 계산 특성을 가질 뿐 아니라 물리적으로 유연한 연결구조를 가질 수 있고 시간적(temporal)으로도 더 유연한 계산특성을 가진다.[4]

1.1 연구의 목표 및 범위

본 연구는 글에서 정서적 정보를 인식해내는 모델을 제시한다. 정서적 정보가 표현되는 양식에는 글 외에도 표정이나 목소리, 몸동작 같은 다양한 것들이 있을 수 있지만, 글은 사람의 언어의 문자 형태라는 점에서 매우 중요한 표현 양식이다. 멀티미디어의 시대임에도 종이에 기록된 전통적인 메모나 편지를 비롯해서 인터넷의 웹 페이지나 전자 우편, 메신저 등 여전히 매우 비중 있는 의사 소통 매체 중의 하나이며 가장 보편적인 지식 저장 매

[3] [Damasio, A. (1994)]

[4] [장병탁, (2005)] 참조.

체이기도 하다. 그리고 지금도 많은 컴퓨터 프로그램들과 로봇들이 글을 토대로 사람과 정보를 주고 받고 있다.

본 연구에서 제시하는 모델은 자연 언어로 작성된 글을 입력으로 받아들인다. 입력된 텍스트는 전처리 과정을 거쳐서 단어 집합 형태로 변환되고 전체 의미 네트워크로부터 추론에 필요한 부분 네트워크를 유도해내기 위한 조건으로 사용된다. 의미 네트워크는 단어들간의 의미적 관계에 대한 정보를 담고 있는데 각각의 단어가 가지는 정서적 정보와 결합되어 어떤 주제나 개념을 나타내는 단어 혹은 단어들의 정서적 의미를 계산하는데 사용되고, 이러한 과정을 통해 계산된 정서적 의미가 모델의 출력이 된다.

1.2 논문의 구성

본 논문에서는 먼저 텍스트 자극과 정서에 관한 심리학적 연구를 살펴보고 이것을 토대로 의미 처리 과정과 정서와의 관계에 대해 2장에서 논의하겠다. 3장에서는 모델의 처리 과정을 구현하는데 사용한 분자컴퓨터에 대해서 간단히 설명하고, 분자컴퓨터가 가지는 장점과 인지과학적 의의에 대해서 논의해보겠다. 텍스트의 전처리 과정에 대해서는 4장에서 살펴보고, 5장에서는 DNA 컴퓨터를 이용한 정서 인식 모델을 소개한다. 6장에서 시뮬레이션에 사용된 자료와 그 결과를 분석하고 토의한 후, 마지막으로 7장에서 본 연구의 내용을 요약하고 문제점 및 향후 과제에 대해 논의해보겠다.

2. 의미 처리 과정과 정서

2.1 Spider Phobia

거미공포증(spider phobia)은 동물공포 중에서 가장 흔한 것으로 아직까지도 정확한 병인은 밝혀지지 않고 있다. 예전에는 과거 생존에 큰 위협이었던 동물들에 대해서 자동적으로 공포감을 가지게 된다는 preparedness hypothesis[5]에 의해 동물공포를 설명하였지만, 최근의 가장 유력한 심리학적인 설명은 거미가 가지는 생김새나 움직임 같은 특정 속성이 혐오감을 불러일으키는 원인이 된다는 것이다.[6]

피험자의 구두 보고에 의존하였던 Seligman의 연구와 달리, Watts와 동료들은 피험자들이 색깔이 칠해진 단어를 선택하는 유사 스트룹 과제(Stroop-like task)를 수행하게 하였다. 그 결과 정상인 피험자들과 달리 거미공포증이 있는 피험자들은 SPIDER라는 거미에 대한 직접적인 단어 이외에도 HAIRY, BRISTLY, CRAWL 같이 의미적으로 거미를 연상시킬 수 있는 단어들에 대해서도 느린 반응 시간을 보였다.

이 결과가 동물공포에 대한 근본적인 설명으로써 어째서 털이 숭숭 난 기어 다니는 작은 동물을 무서워해야 하는지는 설명해주지 않지만, 공포의

---

[5] [Seligman, (1971)]

[6] [Watts et al., (1986); Barker et al., (1997)]

원인이 뱀이나 거미 같은 위협이 될 수 있는 각각의 동물에 있는 것이 아니라 털이 있다거나 다리가 많은 것 같은 그들의 속성에서 비롯된 것이라는 주장을 뒷받침해준다.

두 개의 서로 다른 단어가 의미적으로 깊은 연관을 가지고 있다면 그 두 단어는 정서적으로 비슷한 반응을 불러일으킬 가능성이 높을 것이다. 정서는 즉각적이고 낮은 수준의 인지과정이지만 여기에는 단어 사이의 의미적 관계에 대한 지식이 관여한다.[7]

2.2 정서의 정의

"Everyone knows what an emotion is, until asked to give a definition" [8]

사람들은 정서를 기계적인 계산과는 동떨어진 유기체만의 고유한 특징으로 여기려는 경향이 있다. 사람과 똑같이 걷고 생각하고 사람의 언어로 대화하는 기계가 언젠가는 발명될 것이라 생각하면서도, 기쁨과 슬픔을 느끼고 누군가와 사랑에 빠지는 기계의 존재에 대해서는 부정적인 견해를 가

---

[7] 단어들의 의미적 연관관계와 정서와의 상호 관계는 affective semantic priming에 관한 연구 및 관련된 연구들 참조. [Hermans, (1998)], [Houwer, (2002)]

[8] [Beberly Fehr & James Russel, (1984)] p.464

진다. 직관적으로도 질서정연한 이성에 비해 정서는 언제 어떻게 변할지 예측할 수 없는 비과학적인 대상이다.

그래서인지 정서를 과학의 대상으로 삼고 분석적으로 접근하려는 시도는 역사가 그리 길지 않다. 철학자들은 이성과 감정의 관계를 주인과 종의 관계로 생각하였고, 마음을 과학적 연구 대상으로 삼는 심리학자들조차 초기에는 정서를 사람의 행동을 무질서나 혼돈에 빠뜨리는 힘으로 생각하였고 심지어 일부 학자들은 정서를 쓸모 없고 유해한 것으로 간주하였다.

초기 심리학에서 감정은 언어나 지각, 학습과 같이 심각하게 고려되지 않고 여분의 기능으로 취급되었지만, 정서는 여분으로 취급하기에는 그 영향과 능력이 광범위하며 중요하다. 감정이야말로 사람들이 심리적, 육체적, 지각적으로 경험할 수 있는 사건의 의미와 관계를 다방면으로 연결시켜줄 뿐만 아니라 그것들을 가장 명백하게 평가해서 표현할 수 있는 체계이다.[9]

Antonio Damasio는 뇌에서 감정을 느끼게 해주는 역할을 하는 전두엽이 손상되었을 때, 단순히 감정을 느끼지 못할 뿐 아니라 일상 생활에서도 제대로 일을 계획하거나 실행할 수 없다는 것을 발견했다.[10] 일반적으로 사람들은 감정의 간섭 없이 이성에만 의지해서 의사결정을 한다면 더 객관적이고 올바른 판단을 내릴 수 있을 것이라 생각한다. 하지만 Damasio의 연구는 오히려 예상과 정반대의 결과를 보여준다. 언어적인 능력과 수학적인 능력은 정상임에도 불구하고 전두엽에 손상을 입은 환자는 의사결정에 심각한

[9] [Oatley, K., & Jenkins, J. M. (1996)] pp.95-132.

[10] [Damasio, A. (1994)]

결함을 보였다. 이러한 신경과학적 연구 결과는 최근의 의사결정과 정서에 대한 심리학적 연구[11]와 잘 맞아 떨어지는데, 불확실한 상황하에서의 판단과 의사결정에 관련된 많은 심리적 현상들이 위험을 감정을 통해 추정하는 사람들에 내재된 속성에 대한 이해를 통해서 설명할 수 있다는 것이다.

정서에 대해서 심리학에서 공동으로 받아들여진 하나의 정의는 없다. 과학에 있어 정의란 개념적인 것이라기 보다는 실용적인 것인데 여기서의 실용적인 정의는 정서를 인식하기 위한 것이다. 즉 정서에 관련된 새로운 발견이 이루어진다면 언제든지 바뀔 수 있는 정의들이다.

2.2.1 정서(emotion)와 기분(mood)

우리가 흔히 말하는 감정 또는 느낌은 0.5 ~ 4초 동안에 지나지 않는 일시적인 상태이다. 사람들에게 경험했던 감정들을 계속 기록해놓으라고 하지 않는 한 일상적으로 끊임없이 경험하는 감정들을 기억하고 분류하기란 불가능할 것이다. 그러나 기록된 감정일지라도 몇 분 혹은 몇 시간이 지나면 사람들의 의식에서 사라지기 마련이다. 의식에서 사라진 감정이라 할지라도 언제든지 상상이나 기억을 통해 마치 일기장을 열어보는 것처럼 감정을 다시 경험할 수 있다. 같은 사건에 대해서도 사건을 경험했던 당시와 나중에 다시 떠올릴 때의 감정은 아마 다를지도 모른다.

---

[11] [Loewenstein, (2001)]

이러한 정서는 지속시간이 매우 짧다는 점에서 기분(mood)이나 성격(personality)과는 다르다. 짧게는 몇 분에서 길게는 몇 시간 이상 지속될 수 있는 기분이나 매우 긴 시간 동안 지속되는 성격은 시작과 끝이 매우 모호하기 때문에 원인을 명확히 하기가 어렵다. 감정의 지속시간에 대한 최근의 심리학 실험은 의식에 떠오르지 않는 매우 짧은 시간 동안에만 지속되는 무의식적 정서(unconscious emotion) [12]의 존재에 관한 것이다. 이 연구에 따르면 사람이 판단이나 결정을 내리는 과정에 의식에 떠오르지 않는 역하 수준의 정서가 개입되고 영향을 미친다는 것이다.

일반적으로 짧은 일시적인 감정 상태를 정서(emotion)로 정의하는데, 본 연구에서 다루고자 하는 정서도 SPIDER라는 단어를 봤을 때 짧은 순간 떠올리는 혐오감 같은 제한적인 감정이다. 즉 정서는 주어진 단어에 대해서 무의식적인 병렬 처리 과정을 통해 유도되는 극히 짧은 시간의 감정 상태로 정의된다.

2.2.2 기본 정서 (Basic Emotion)

그렇다면 정서는 도대체 무엇이고 왜 생겨난 것일까? 정서의 존재에 대한 이런 근본적인 질문에 대해서, 기능적 관점에서 내놓을 수 있는 최선의

---

[12] [Winkielman, (2004); Berridge & Winkielman, (2003)] 참조.

무의식적인 정서의 존재는 감정은 의식에 의해 인식(percept)되어야 비로소 존재하게 된다는 기존의 전통적인 생각에 반하는 것이었다.

답변은 아마도 유기체가 주어진 환경에 적응한 결과라는 진화론적 설명일 것이다. 유기적 협동을 유지시켜 나가는 것, 물리적 위협으로부터 벗어나는 것 같은 생존에 직결되는 문제들을 해결하기 위한 방안으로써 정서라는 기능이 발달하게 되었다는 것이다. 이러한 생물학적 가정을 바탕으로 유기체의 생존에 필요한 전략과 직접적으로 결부되는 몇 가지 정서를 기본 정서(basic emotion)로 정의하였다.[13]

이러한 기본 정서에 바탕을 둔 접근법은 신경학적 두뇌 연구와 밀접한 관련이 있다. 포유류의 두뇌는 단일화된 감정 시스템을 가지고 있는 것이 아니라 각각의 기본 정서에 대한 체계가 잘 분화되어 있는데, 이러한 체계가 학습과 문화적 경험들을 통해서 사회적으로 구성된 정서의 모태가 된다고 본다. 모든 정서는 피질상의 강한 흥분을 촉발하지만 피질이 감정을 불러일으키는데 필수적인 것은 아니다. 단지 인지적 정보가 감정적 흥분과 관계된다는 것을 의미할 뿐이다.

감정 시스템들은 변연계 뉴런들의 전기적 신호들뿐만 아니라 다양한 종류의 신경전달물질과 신경조절물질들의 상호작용으로 구성된다. 가령 분노에 관한 시스템은 글루타민(glutamate) 같은 흥분성 아미노산 신경전달물질에 의해 조절되고, 도파민(dopamine)은 긍정적인 감정과 깊은 관련이 있다. 이러한 화학적인 상호작용 체계는 시냅스 수준에서 국지적으로도 일어나지

---

[13] [Oatley, K., & Johnson-Laird, P. N. (1989)], [Oatley, K., & Jenkins, J. M. (1996)] 참조.

만 혈관을 타고 수용체로 이동하는 호르몬처럼 오랜 시간에 걸쳐 광범위한 영역에 영향을 준다.

많은 수의 정서를 몇 개의 기본 범주로 나누어 인식하는 접근 방법은 실용적인 관점에서뿐만 아니라 이처럼 생물학적으로도 타당성이 있다. 기본 정서의 종류와 개수에 대해서는 연구자의 관점과 목적에 따라 다양한 정의가 있는데, 여기서는 일반적으로 받아들여지는 기쁨(happy), 슬픔(sad), 분노(angry), 공포(fear), 혐오(disgust), 놀람(surprising)의 6가지 정서를 사용한 기본정서 정의를 사용한다.

2.2.3 정서와 일반지식(Commonsense)

정서를 다른 정서와 어떻게 구분하느냐는 문제에 대해 신경과학적 발견에 바탕을 둔 기본정서를 이용한 설명은 유력한 해답중의 하나이다. 이러한 관점은 특정 정서 범주가 매우 공통적이며 심지어 보편적인 것임을 제시하는 여러 증거들에 잘 부합된다. 정서의 생물학적 기초에 대해서 반대하는 사람은 아마 없을 것이다.

이러한 생물학적 기질과 더불어 정서에는 문화적인 결정 인자들이 존재한다. 사람들의 정서는 문화마다 차이가 있다. 서로 다른 풍습의 차이로 정서적 오해를 불러일으키는 경우를 우리는 주변에서 쉽게 찾아볼 수 있다. 하지만 문화적 차이가 생물학적 기본 정서의 한계를 뛰어넘어 정서 자체를 바꿀 수 있는 것은 아니다. 다른 문화권의 사람이라도 상대가 정상적인 사람이라면 화가 난 이유는 알 수 없어도 화가 났다는 사실은 금새 알아챌 수

있다. 만약 정서가 완전히 문화에 좌우되는 것이라면 서로 다른 문화권의 사람들끼리의 정서적 의사 소통은 거의 불가능하다. 마치 흑백으로밖에 보이지 않는 사람에게 색깔을 구분하라고 요구하는 것과 같을 것이다.

2.3 정서 인식에 대한 인공지능 연구

인공지능 알고리즘을 이용해 텍스트에서 정서를 인식하려는 연구는 이미 다양하게 시도되어왔다. 본 모델의 특징은 대용량의 실제 세계에 대한 지식을 기반으로 논리적인 추론을 사용하면서도 화학적 병렬처리 알고리즘을 이용하여 텍스트를 통계적으로 접근한다는데 있다. 이번 장에서는 기존의 텍스트 처리 알고리즘과 비교하여 장단점에 대해서만 간략히 언급하고 5장에서 본 모델에 대해 구체적으로 설명하겠다.

핵심어 검출 방법(Keyword spotting)은 가장 단순하지만 아마도 널리 사용되는 방법일 것이다. 이 방법은 단지 특정 단어의 존재 유무에 따라 텍스트의 정서를 판단한다. 예를 들어 “today was a happy day”에서 “happy”가 행복함을 나타내므로 주어진 문장을 행복함으로 분류하는 것이다. 이러한 방법은 쉽게 사용할 수 있지만, 부정문은 처리하기 곤란하고 단어의 표면적 의미만을 사용한다는 것이 한계점이다.

단어 연관성(Lexical affinity)를 사용한 방법은 단어마다 특정 정서에 대한 확률값을 지정하는 것이다. 가령 “accident”는 자동차 사고나 부상 등을 떠올리기 때문에 75%의 부정적 의미를 가진다고 정의하는 것이다. 이러한

확률값은 언어자료(corpus)를 이용해서 학습시킬 수도 있고 핵심어 검출 방법에 비해 높은 정확도를 보인다. 하지만 역시 부정문은 처리가 불가능하고, 학습시킨 데이터에 따라 전혀 다른 성격의 글에 대해서는 제대로 동작하지 않는 경우가 생긴다. 따라서 재사용 가능한 영역 독립적인(domain-independent) 모델을 만드는 것이 매우 어렵다.

세 번째는 통계적인 언어 처리 방법이다. 방대한 언어자료(corpus)를 이용해 단어의 빈도와 분포를 이용하는 방법이다. 단점은 신뢰할만한 출력 값을 위해서는 많은 양의 입력 텍스트가 필요하다는 점이다. 문단 수준의 몇 백 단어 이상의 긴 글에서는 안정적으로 동작하지만 한 문장이나 몇 단어의 수준에서는 정확도가 떨어진다. 그리고 통계 데이터는 의미론적으로 강하지 못하다는 것이 또 다른 단점이다. 부정문이나 엉뚱한 단어와의 조합에 대해 개별적으로 예외적인 처리를 해줄 수 없기 때문이다.

네 번째는 Schank와 Dyer의 텍스트 이해 모델이다. 이 모델은 사람의 욕구, 욕망, 목표에 대한 미리 준비된 조건들을 가지고 주어진 텍스트를 이것들에 비추어 분석한다. 텍스트에 대해서 깊은 이해를 바탕으로 분석을 할 수 있는 큰 장점이 있지만 스크립트에 기반한 기호주의(symbolic model) 시스템의 한계로 인해 준비된 텍스트 이외의 임의의 텍스트로의 확장이 사실상 불가능한 단점이 있다.

본 연구에서 제시하는 모델은 핵심어 검출 방법처럼 표면적 의미에 의존하지 않고 각각의 단어들의 내포된 의미를 확률적으로 사용한다. 통계적인 방법을 사용하여 모델의 확장이 용이하며, 입력벡터를 생성하는 방법이

통계적인 것이 아니라 입력벡터가 처리되는 과정이 통계적이기(high fluctuation) 때문에 통계적인 모델임에도 불구하고 몇 단어의 적은 수의 입력에 대해서도 정확도가 크게 손상되지 않는다. 그리고 특정 문제 영역에 관련된 언어자료가 아닌 일반적인 지식을 담고 있는 데이터베이스(Commonsense DB)를 모델의 추론에 사용함으로써 영역 독립적인 모델이 될 수 있도록 하였다. 마지막으로 단어를 철자 수준에서 직접 사용하지 않고 의미 전처리 과정을 거친 후 보다 개념적인 수준에서 단어를 사용함으로써 추론 과정에서 오류가 덜 발생하도록 고려하였다.

3. 인지과정과 분자 컴퓨터

3.1 분자 컴퓨터(Molecular Computer)

분자 컴퓨터는 1985년 Conrad에 의해 고안되었으며, 1994년 Adleman에 의해 DNA 분자를 사용하여 시험관내에서 처음으로 구현되었다.[14] 생물체의 세포 속에 존재하는 DNA가 AGCT의 4가지의 염기로 이루어져 있고 염기들 간의 상보 결합에 의해 유전 정보가 저장되고 복제됨에 착안하여 DNA 분자를 이용해 Hamiltonian 문제를 푸는데 성공하였다. DNA 컴퓨터가 주목 받

[14] [Adleman, (1994)]

는 이유는 그 초병렬적인 정보처리 능력에 있다. 단 3g의 물에 약 10^{20}개의 DNA 분자를 담을 수가 있는데 화학 반응에 의해 각각의 분자들이 담고 있는 정보들이 거의 동시에 한번에 계산되는 특징을 가진다. 기존의 실리콘 컴퓨터가 한 번에 하나씩 계산을 처리할 수 있는 반면 화학 반응을 이용하는 DNA 컴퓨터는 하나의 계산이 일어나는 속도는 약간 느리지만 한 번에 계산할 수 있는 용량이 비교할 수 없을 정도로 크다. 분자 컴퓨터의 이러한 초병렬성은 나노 수준의 미시 시계에서 일어나는 분자들의 인식(molecular recognition) 특성 때문인데, 이러한 분자인식을 통한 자기조립(self-assembly) 알고리즘과 시험관내의 진화 알고리즘[15]을 이용하여 사람의 인지 과정을 설명하려는 연구[16]들이 이루어지고 있다.

3.2 마음의 모델링

1930년경 튜링기계가 인지과정을 형식체계(formal system)로 환원하여 자동화하려는 시도로부터 시작된 마음의 기계화 작업은 인공신경망의 발명과 함께 비약적으로 발전하게 된다. 인공신경망은 범용튜링기계와 같은 수준의 계산 능력을 가지고 있고, 두뇌에서 뉴런들이 작동하는 방식처럼 분산 처리

---

[15] PLM(Probabilistic Library Model)에 의한 학습과 추론 알고리즘은 [Zhang, B. T., & Jang, H. Y., (2004); (2005)] 참조.

[16] [김지수, (2005)], [이은석, (2003)] 참조.

와 병렬 처리 특성을 가지고 있어서 연구자들의 폭넓은 지지를 받고 있다. 사람의 마음에 대한 기존의 기호주의 접근방식에 비해 인공신경망을 이용한 연결주의 접근방식은 앞서 언급한 장점들을 가지지만 두뇌 뉴런의 전기적 신호 전달만을 모사한다는 점에서 한계가 있다.

3.2.1 생화학적 신경망

사람의 몸은 약 10^{12}개의 세포들로 이루어져있고 각 세포들은 보통 10^4가지 종류의 총 10^9개의 단백질 분자들을 포함하고 있다. 각각의 세포막 외부에는 특정 단백질 분자들을 받아들이는 수용기(receptor)가 존재하는데, 단백질 분자에 의한 신호를 수용기가 받아들이면 신호전달 단백질(signal-transduction proteins)이 활성화되어서 효소에 의해 영향을 받는 특정 화학 반응을 세포 내에서 촉진하게 된다. 이러한 세포간의 단백질 분자들에 의한 상호작용은 세포를 분열시키거나, 죽이거나, 다른 신호 단백질을 생성하게 하는 등 다양한 결과를 불러일으킨다. 이러한 생화학적 경로(biochemical pathway; signal transduction pathway)에 포함된 세포들은 특정 신호 단백질에 의해 동시에 초병렬적으로 이러한 일종의 계산 과정을 처리하게 된다. 개개의 세포들을 특정 신호들을 처리하도록 설정된 작은 처리 단위(마치 뉴런처럼)로 생각한다면 이러한 생화학적 경로에 의한 네트워크도 신호를 주고받아 정보를 처리하는 시스템으로 생각할 수 있다.[17]

---

[17] [장병탁, (2005)], [Thagard, (2002)] 참조.

이러한 신경화학적 네트워크는 기존의 인공신경망에 비해 세 가지 면에서 질적으로 큰 차이를 보인다. 먼저 신경화학적 네트워크는 처리 단위의 내부적 처리 용량이 훨씬 크다. 기존의 뉴런이 단순하게 입력을 합하여 커널 함수를 거쳐서 출력을 내놓는 것이 비해, 세포는 다양한 신호 단백질 분자에 대해 세포 내 특정 화학 반응을 조절하거나, 인접한 세포에 전기적 신호를 전달하거나, 호르몬 같은 신호 단백질을 이용해서 멀리 떨어진 다른 세포들에 화학적 신호를 전달하는 등 다양한 출력을 내놓는다. 단지 활성화된 정도 혹은 돌출(spike) 패턴만을 저장하고 처리할 수 있는 인공 뉴런에 비해 실제 뉴런 세포는 훨씬 높은 표현력을 가진다. 둘째는 신경전달물질의 생화학적 경로에 의한 공간적 특성이다. 기존의 신경망에서 뉴런들 사이의 연결은 층위(layer)를 이루어 미리 결정되어 있어야 하지만, 생화학적 네트워크에서는 기능마다 처리 경로(혹은 네트워크)가 독립적으로 존재하고 혈관을 매체로 하기 때문에 처리 단위 사이의 연결 설정이 훨씬 자유롭다. 그리고 처리 경로에 관여하는 임의의 연결된 세포 혹은 시스템들이 정보를 공유할 수 있는 메커니즘을 제공하기 때문에 몸이나 두뇌 전체의 일관적인 동작 양식을 조절할 수 있게 한다. 이러한 역할은 주로 호르몬 같은 신경조절물질에 의해 이루어지는데 두뇌 뉴런의 동작은 인접한 뉴런들로부터 전달된 전기적 신호뿐만 아니라 이러한 생화학적 요소들까지 고려된 매우 복잡한 처리 특성을 가진다. 셋째는 신호 단백질에 의한 전달 과정은 전체 처리 경로에서 비동기적(asynchronous) 성격을 가진다는 것이다. 인공신경망에서의 병렬처리는 입력에 대한 출력 값을 계산하기 위해 각 뉴런들간의 동기화된

계산 과정을 필요로 한다. 그러나 생화학적 신호의 경우 시냅스에서 처리되는 일반적인 신경전달물질이 있는 반면에 몇 시간이나 몇 일 이상 지속되는 호르몬도 존재한다. 그리고 특정 호르몬은 수용기가 있는 뉴런 세포의 발화 속도에 영향을 준다. 예를 들면 에스트로겐은 도파민과 세로토닌의 분비량에 영향을 줌으로써 뉴런의 발화 속도를 늦추거나 높일 수 있다. 즉, 신호 단백질에 의해 뉴런 네트워크의 비동기적 처리 특성을 바꿔줄 수 있는 것이다.

3.2.2 정서와 인지와 신경화학

사람의 내적 처리과정에 전기적 신호 외에 화학적 신호가 중요한 역할을 한다는 관점에서 언어나 논리적 사고 같은 상위 인지 과정에 직접 접근하는 것은 신경화학적 연구 사례를 찾기 힘들고 오히려 이러한 인지 과정은 기호주의(symbolism)를 통해 더 잘 설명되는 것처럼 보인다. 하지만 이러한 화학적 신호는 2장에서 살펴보았듯이 정서와 매우 밀접한 관련이 있으며 의사 결정이나 판단 같은 상위 인지 과정에도 정서가 중요한 역할을 담당한다는 실험적 증거들을 앞서 소개하였다. 따라서 사람의 인지과정을 이해하기 위해서는 생화학적인 수준에서 뉴런 세포들의 동작과 동역학(dynamics)적인 특징의 이해는 필수적이라 말할 수 있다.

이러한 생물학적이고 화학적인 생명체의 동적인 특성을 도입하여 보다 지능적인 시스템을 구현하려는 연구[18]가 다양하게 시도되고 있는데, 분자컴퓨터를 이용하려는 시도는 신경화학적 신경망의 초병렬성, 공간적 특성, 시간적 특성을 만족시키는 새로운 패러다임의 접근 방법이다. 본 연구에서 제시하는 분자컴퓨터 모델은 실제 세포들의 단백질 상호작용 네트워크처럼 정교한 처리 과정을 보여주지는 못하지만 화학 반응에 기반한 초병렬성과 더불어 분자인식(molecular recognition)을 통한 임의적인 연결과 비동기적 자기조립과정을 사용하여 정서를 인식해내는 과정을 보여준다.

3.3 일반지식(Common sense)에 의한 추론

3.3.1 일반지식(Common sense) 컴퓨터

오늘날 컴퓨터 프로그램들은 여러 분야에서 효율적이고 정확하게 맡은 역할을 수행하고 있다. 문자나 얼굴을 인식하고, 공장에서 복잡한 기계를 조립하고, 자동차나 비행기도 조종한다. 사람보다 체스를 잘 두는 프로그램도

[18] 지능 시스템에 정서 및 자연적 신호전달체계의 생물학적, 화학적 특성을 이용하려는 대표적인 연구는 [Sloman, (2005)] 및 [Brooks, (2002)] 참조.

개발되었고, 사람의 질병을 진단해내기도 한다. 하지만 글을 읽거나 집을 청소하거나 아기를 돌보는 지능적인 기계는 아직 세상 어디에도 없다.[19]

글에 담긴 정서를 인식해내는 것도 글을 읽을 수 있는 사람이라면 힘들이지 않고 자연스럽게 할 수 있는 쉬운 일이지만 컴퓨터 프로그램에게는 흉내도 낼 수 없는 매우 어려운 과제이다. 과연 양자간의 차이가 단지 10^{11}개의 뉴런 세포와 10^{14}개의 시냅스 탓으로 돌려버릴 수 있는 것일까? 더 좋은 처리 속도와 저장 용량을 제공한다고 해서 지금의 프로그램이 언젠가 그러한 일을 자연스럽게 해낼 수 있게 되는 것일까? 이러한 질문들에 대해 Minsky를 비롯한 많은 학자들은 두뇌의 용량이나 속도가 월등하기 때문인 것은 아니라고 주장한다.

Tomas Landauer에 따르면 사람의 두뇌가 일생 동안 학습하는 정보는 $4*10^9$ 비트 정도에 불과하다. 단순하게 지식의 양만 본다면 사람이 평생 배워야 할 분량을 컴퓨터는 씨디 한 장에 담아서 단 몇 시간 만에 배울 수 있다. 그러나 현재의 컴퓨터 프로그램에 쌓여진 지식은 특정 문제를 풀기 위해 전문화된 지식인 반면, 사람의 두뇌에 쌓여진 지식은 특정 문제 영역에 종속된 지식이 아닌 일반적인 지식이 대부분이다. 이러한 지식이 바로 상식(common sense)인데, 이것을 이용해서 다양한 문제를 해결할 수 있는 능력이 사람과 컴퓨터의 가장 큰 차이점이다. 사람이 정서를 처리하는 과정에도 이러한 일반적인 지식이 활용된다. 정서가 보편성을 가지지만 문화적인 요인

---

[19] [Minsky, (forthcoming)] 참조.

에 따라 사람들의 정서 인식과 표현 방법이 현격한 차이를 보이는 현상도 이러한 주장의 근거가 된다.

Minsky는 일반지식(commonsense)을 이용하는 범용적인 문제 해석기(universal solver)의 구현을 위해서는 세 가지 요소가 필요하다고 주장하는데, 지식에 대한 범용적인 정합 알고리즘과 그 지식들의 연결 알고리즘 그리고 목표 값에 대한 오차를 줄이는 범용적인 추론 알고리즘이 그것들이다. 분자 컴퓨터는 단순하게 많은 계산을 동시에 수행할 수 있다는 장점 외에도 이러한 요구 조건에 상당 부분 부합하는 계산 특성을 지니고 있다.

3.3.2 분자컴퓨터와 일반지식(Commonsense) 추론

분자 컴퓨터의 초병렬적 계산 특성을 가능하게 하는 것은 나노 수준의 미시적 세계에서 분자들이 서로를 인식하는(molecular recognition) 메커니즘이다. 우리 몸 안에서 특정 신호 단백질을 특정한 수용기에서만 인식하는 것이나, 특정 세균이나 바이러스에 대한 면역 체계의 항원-항체 반응이 쉽게 찾아볼 수 있는 예이다. DNA는 이러한 분자 인식 메커니즘을 이용해 자연계에서 실제로 유전 정보를 저장하는데 사용되는 매체이고, AGCT의 염기간의 상보 결합을 이용하여 구현한 것이 분자컴퓨터의 한 종류인 DNA 컴퓨터이다.

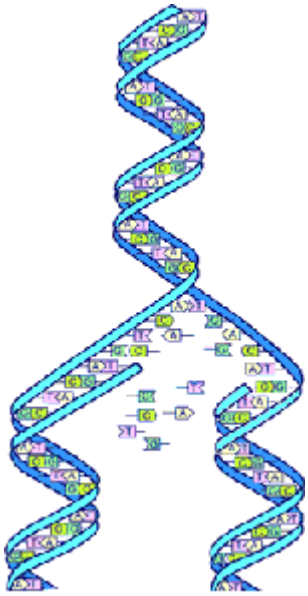

그림 1 DNA의 분자 인식에 의한 이중나선구조

만약 이러한 분자 수준에서의 정합 알고리즘을 바탕으로 특정 나노 분자에 정보를 저장하고 인출하는 기억장치가 있다면 이 기억장치는 자연스러운 연관메모리가 된다. 기존의 튜링머신(Turing Machine)은 주소(address)에 기반한 기억장치이고 이러한 장치에서 주소가 아닌 값(value)에 의한 참조를 하기 위해서는 매우 높은 비용을 감당해야 한다. 예를 들어, N개의 레코드를 가지는 주소 메모리 체계에서 “molecule”이라는 문자열이 포함된 모든 레코드를 검색하기 위해서는 N개의 모든 레코드에 대해서 순차적으로 값을 검사(exhaustive search)해야 한다. 물론 인덱스 구조를 따로 유지해서 검색을 효율적으로 할 수는 있겠지만, 인덱스 되지 않은 변수에 대해서는 여전히 높

은 비용을 지불해야 하고 모든 경우의 수에 대해서 인덱스를 생성하는 것은 속도와 용량 모두 큰 부담이 될 것이다.

이러한 주소기반 기억장치의 한계점이 기존의 컴퓨터 프로그램을 특정 문제에 종속적인 지식만을 유지하고 검색할 수 있게 하는 요인중의 하나라고도 생각해 볼 수 있다. 일반지식(commonsense)을 문제 해결에 이용하려면 다양한 영역에 걸친 많은 정보들 중에서 관련된 정보만을 검색하고 수집해야 하는데 주소기반의 기억장치보다는 연관메모리에 의한 프로그램이 이러한 기능을 수행하는데 본질적으로 유리할 것이다.

분자인식에서 비롯된 분자 컴퓨터의 또 다른 특징은 자기조립(self-assembly, self-organizing)이다. 분자들간에 서로 결합할 수 있는 부분이 분자간의 결합이 이루어진 후에도 계속 남아 있도록 구조를 설계하면 환경이 허락하는 한 분자들이 계속 스스로 연쇄적인 결합을 이루게 된다. 이러한 방식에서는 분자들이 서로 자기조립되어 나가는 과정이 계산이며 그 결과 만들어진 나노 구조가 계산 결과가 된다. 자기조립은 주어진 문제를 해결하기 위해 기억장치에서 검색된 지식의 단편(fragment)들을 서로 어떻게 연결시킬지에 대한 훌륭한 해결 방안이 될 수 있다.

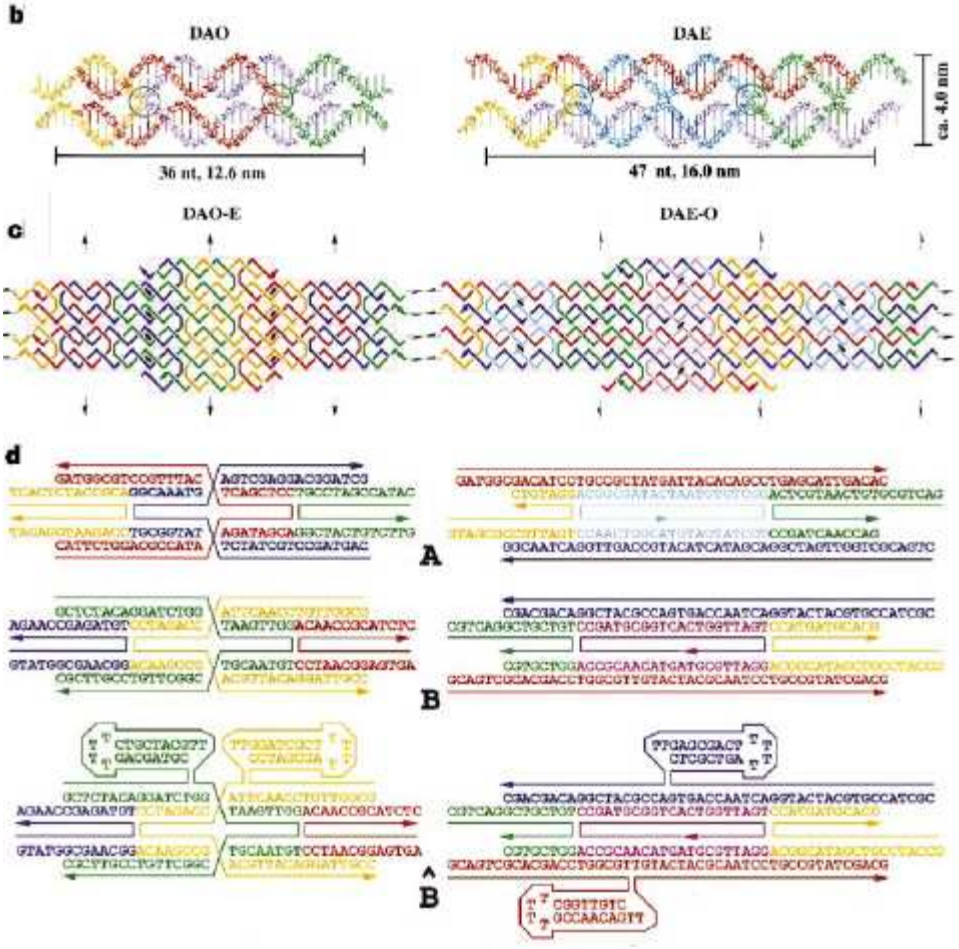

그림 2 DNA 분자들의 2차원 자기조립의 예

분자들 사이의 친화도(affinity)는 분자인식 메커니즘의 기본이 되지만 생성된 분자의 구조에 의해 이미 결정되어 버렸기 때문에, DNA 사슬을 조작하는 나노 기계의 힘을 빌리지 않는 한 쉽게 바꿀 수 없는 특성이다. 하지만 PLM(Probabilistic Library Model)은 많은 수의 분자를 이용하여 데이터 변수들의 경험적인 결합 확률을 표현하고 분자컴퓨터의 초병렬성을 사용하여 시험관내에서 확률적인 추론과 학습을 수행하는 범용적인 알고리즘을 제

시해준다. 분자컴퓨터의 연관메모리와 자기조립 특성을 분자들의 배열 구조를 바꾸지 않고도 시험관내의 분자들의 분포 수를 이용하여 통제할 수 있는 것이다. 쉽게 설명하면 분자간의 친화도(affinity)가 다소 떨어지는 경우에라도 많은 수의 분자를 넣어주면 상대적으로 더 많이 반응하게 되고, 친화도가 높은 분자들일지라도 시험관내에 적은 수만 존재한다면 결과 구조물의 생성에 별로 기여를 하지 못하게 될 것이다.

3.4 의미 네트워크(semantic network)

PLM 모델에서 분자들이 단어를 나타내고 분자들 사이의 관계가 확률적으로 표현된다면 이것은 단어와 단어 사이의 연합 관계를 나타내는 의미 네트워크로 볼 수 있다. 의미 네트워크(semantic network)는 의미 정보를 저장하기 위한 수단으로 Collins와 Loftus(1975)에 의해 제시되었는데 노드 사이의 임의의 관계를 표현할 수 있다는 점에서 사람의 연상기억이나 점화, 간섭을 설명하는데 주로 사용되었다. 그래프 구조를 어떻게 해석하느냐에 따라 분류체계(taxonomy)나 유사성 같은 정보를 저장하고 인출할 수 있는데, 여기에서 사용하는 의미 네트워크에서는 서로를 연상시키는 관계에 있는 단어들을 저장하고 인출하는 구조로 생각해 볼 수 있다.[20] 가령 '나무' 같은

[20] Semantic network from word association, [Griffiths & Steyvers, (2002)] 참조.

특정 단어를 정해주고 관련된 단어들을 자유롭게 연상해보라고 하면 '나뭇잎', '뿌리', '장롱', '푸른색' 같은 단어들을 줄줄이 늘어놓을 수 있을 것이다.

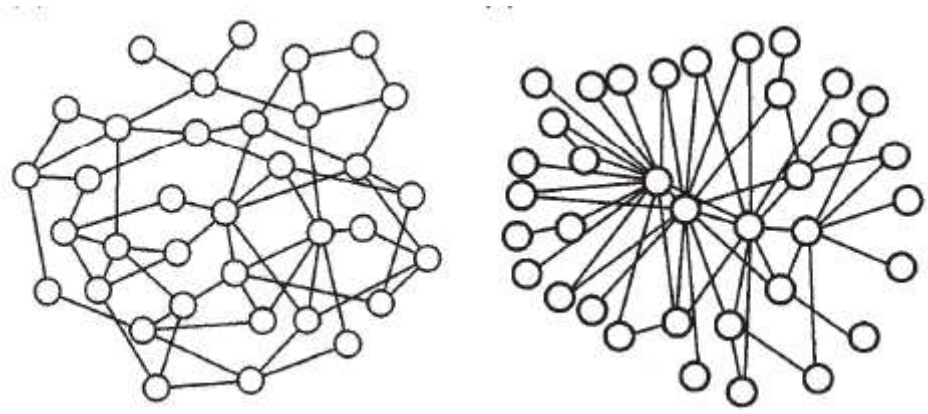

그림 3 Scale-free 의미네트워크의 예

자연 언어를 대상으로 하는 이러한 의미 네트워크가 가지는 주목할만한 특징은 척도 없는 네트워크(scale-free network) 구조가 나타난다는 것이다. 척도 없는(Scale-free) 그래프는 노드의 차수(degree)를 노드의 연결선(edge)로 정의할 때, 노드의 차수와 특정 차수를 가지는 노드의 존재 확률이 파워 분포(power-law distribution)로 나타나는 경우를 말한다. 자연 발생적인(self-organized) 그래프 구조에서 이러한 현상은 피할 수 없는데, 새로운 노드가 그래프에 연결될 때 전체 그래프 구조를 고려하여 연결될 위치를 결정하지 못하고 지역적인 정보에 의존해 결정을 내리기 때문이다. Word association norms[21], WordNet[22], Roget's Thesaurus[23] 같은 언어에 관한 네트워크뿐만 아니

[21] http://www.usf.edu/FreeAssociation

라, 운송 네트워크, 사회적 네트워크, 사업적 네트워크, 생명체 내에서의 신호 체계 네트워크 등에서도 이러한 특징이 발견된다.[24] 그림3 에서 왼쪽 그래프는 일반적인 네트워크이고 오른쪽 그래프는 척도 없는 네트워크인데, 오른쪽 그래프에서는 허브(hub) 역할을 하는 노드와 함께 노드들의 클러스터(cluster)가 관찰되고 임의의 두 노드 간의 평균 거리가 훨씬 짧은 특징이 있다.

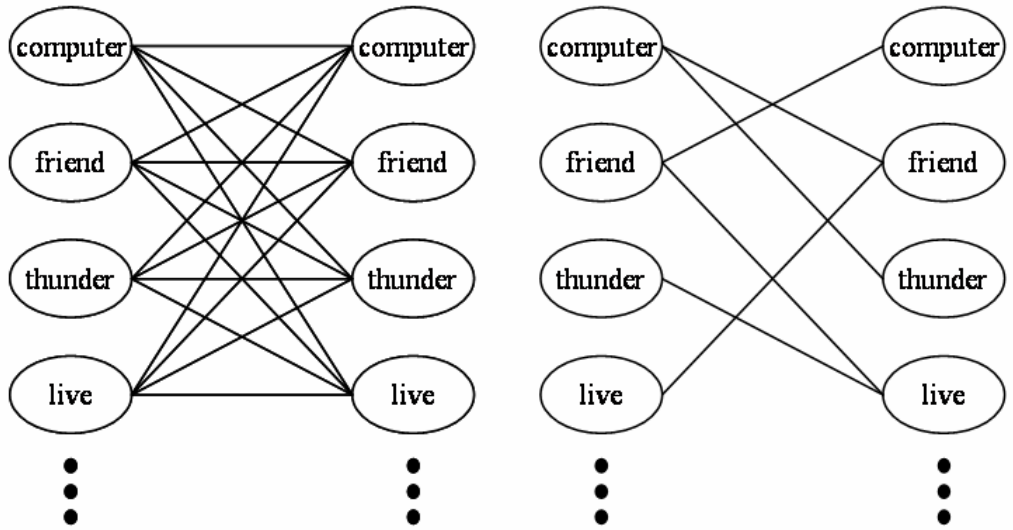

그림 4 단어와 단어 사이의 bi-partite Network

단어들간의 관계는 실제 사람들의 언어 구사에 관한 자료를 이용하여 학습되어 일반적인 지식을 나타내도록 조절된다. 그림 4에서 왼쪽 네트워크

---

[22] http://wordnet.princeton.edu

[23] http://www.gutenberg.org/etext/10681

[24] [Barabasi, (1999)]

는 학습이 이루어지지 않은 초기 상태의 PLM 라이브러리 분포이다. 학습 과정을 통해 의미적으로 관련이 있는 단어들 사이의 연관 강도는 강화되고 그렇지 않은 경우에는 강도가 약화되어 왼쪽 네트워크 형태로 변화하게 되고, 이러한 구조에서 단어의 의미는 연관된 다른 단어들과의 연결 패턴과 강도에 의해 정의된다.

자연언어 의미론(semantics)에 대한 네트워크 기반 접근 방법은 의미를 학습하여 지식을 확장시키고 메모리에서 의미를 검색하는 과정에 대해 앞서 살펴본 것처럼 중요한 함축성을 가진다.[25] 물론 이러한 확률적인 접근 방법으로 본질적인 의미론에 대해 논하는 것은 불가능한 일이고, 오로지 단어들의 연관 관계들로부터 의미 구조의 깊숙한 측면까지 유도해낼 수 있으리라고는 아무도 생각하지 않는다. 대신 단어의 의미와 연관기억의 관계에 주목하여 네트워크 구조를 통해 의미적인 정보가 저장되고 검색되는 과정을 모델링하여 단어의 정서적 의미를 추론해내고자 하는 것이 본 연구의 주된 관심사이다.

---

[25] [Steyvers & Tenenbaum, (2005)]

4. 텍스트 전처리 과정

모델에서 처리되기 위해 텍스트 데이터는 전처리 과정을 거쳐서 벡터 형태로 변환된다. 하지만 단어를 그대로 벡터로 변환하는 것은 변환 과정에서 텍스트의 의미가 손실되는 경우가 생긴다. 간단한 예를 들면, “Mary has gone to New York.”란 문장을 단순하게 단어들로만 분리하면 [‘Mary’, ‘has’, ‘gone’, ‘to’, ‘New’, ‘York’]가 되는데, 철자 수준의 단어 처리는 ‘New York’을 ‘New’와 ‘York’으로 나누어서 처리하게 되어 ‘New York’가 가지는 원래 의미가 소실될 뿐만 아니라 ‘New’, ‘York’라는 관계없는 단어들이 생성되어 추론과정에 관여하게 되므로 심각한 오류를 유발시킬 수 있다. 따라서 최소 단위의 의미를 해치지 않는 범위에서 텍스트를 변환하는 알고리즘을 사용해야 한다. 여기에서는 철자 수준의 단어를 그대로 사용하지 않고 전처리 과정을 거쳐서 조금 더 개념적인 수준의 단어로 변환하여 사용한다. 이러한 전처리 과정의 사용은 물론 완벽한 의미 단위를 추출하는 것과는 거리가 멀지만 가공하지 않은 단어를 직접 사용하는 것보다는 훨씬 견고한 접근방법이다.

4.1 의미 분석기

의미적 최소 단위를 추출하기 위해서는 최소한 주어와 동사가 포함된 의미적 사건 구조에 대한 처리를 포함해야 한다. 여기서는 피상적인(shallow)

수준이지만 이러한 의미 구조에 대한 처리를 이용해 텍스트에서 추출된 개념을 사용하려 한다. 만약 “Mary has gone to New York.” 문장을 GO($person,$place)라는 구조를 이용해서 분석을 한다면 뒤에 오는 ‘New York’이 PLACE에 해당한다는 것을 알 수 있기 때문에 ‘New’와 ‘York’로 분리하지 않고 하나의 처리 단위로 인식할 수 있을 것이다. 의미 분석 과정은 이러한 구조를 문법 분석 결과에 반복적으로 적용시킴으로써 이루어진다.

본 연구에서는 이러한 작업에 Hugo Liu와 Push Singh의 MontyLingua[26] 의미 분석기를 사용한다. 이 분석기는 Jackendoff의 LCS(Lexical Conceptual Structure)에 원리적 바탕을 두었는데 같은 LCS에 바탕을 둔 Dorr의 UNITRAN과는 달리 이벤트 구조를 파싱된 문법 구조에 직접 적용시켜서 의미를 분석한다. Link Grammar Parser of English에 의해 분석된 정보를 바탕으로 텍스트에서 POS tag를 참조하여 직접 이벤트 구조로 연역하는데, 간편하고 문법 구조를 그대로 살려서 사용할 수 있다는 장점이 있는 대신 문법 구조에 의지하기 때문에 하나의 의미에 대한 여러 가지 표현을 처리하는 데에 큰 어려움이 있다. 가령 “Mary can run.”과 “Mary is able to run.”은 문법적 구조가 다르지만 의미상 같은 문장이다. 물론 ‘be able to’를 한 묶음으로 분석은 가능하겠지만 두 가지 경우를 하나의 의미 단위로 분석해내기 위해서는 미리 동의어에 대한 정보를 가지고 있어야 가능할 것이다. 문장의 구조가 완전히 변화하는 경우에는 더욱 이러한 작업이 어려워진다.

---

[26] [Liu, H. (2004)] MontyLingua 프로젝트 웹페이지 참조.

4.2 단어 주머니 (Bag of word)

전처리 과정을 거친 추상적인 수준의 단어는 추론 엔진에 의해 직접 조작되는 기호(symbol)의 역할을 한다. 기호는 철자수준의 '단어'와는 일대다 대응을 가지는 개념적인 수준의 '단어'인데 앞으로 모델에 관련되어 사용된 '단어'는 모두 개념적인 수준의 '단어'를 지칭하는 말이다.

전처리 과정에서는 하나의 동사-논항(verb-arguments) 구조에 대해 하나의 개념적 수준의 단어 집합을 생성한다. 하나의 문장에서도 여러 개의 동사-논항 구조가 발견된다면 각 구조 별로 단어 집합들을 생성한다. 다음 문장을 분석해서 단어 집합을 생성하는 과정을 살펴보자.

"Last weekend, I went to Ken and Mary's wedding in San Francisco."

먼저 구문 분석을 하고 POS tag를 붙인 결과는 다음과 같다.

(S (NP (JJ Last) (NN weekend)) (, ,) (NP (PRP I)) (VP (VBD went) (PP (TO to) (NP (NP (NNP Ken) (CC and) (NNP Mary) (POS 's) (NN wedding)) (PP (IN in) (NP (NNP San) (NNP Francisco)))))) (. .))

문법적 분석(syntactic parsing) 결과에 사건 구조를 반영하여 동사-논항 구조를 분석한다.

[[['go', ['past_tense']], ['I', []], ["to Ken and Mary 's wedding", ['prep=to']], ['in San Francisco', ['prep=in']]]]

필요 없는 분석 정보는 제거하여 원래 텍스트만 남긴다. (Stemming)

[['go', 'I', "Ken and Mary 's wedding", 'San Francisco']]

주어진 문장의 단어 집합은

{ go, I, Ken and Mary’s wedding, San Francisco }

가 되고, 이것이 모델의 입력이 된다.

4.3 한계점

피상적인 수준의 의미 분석기를 이용한 전처리 과정이기 때문에 텍스트에 대한 깊은 이해를 하지 못하는 데에서 오는 한계점이 있다. 첫 번째는 앞서 언급했듯이 하나의 의미를 표현하는 문장에는 여러 개가 있을 수 있는데 이것을 하나의 같은 의미로 유도할 수 없다. 예를 들어 “John met Mary at the movies”와 “John and Mary met at the movies”는 의미상 거의 같은 문장이지만 이벤트 분석을 문장 구조에 의존하는 이러한 방식의 의미 분석기에서는 같은 구조로 유도하지 못한다.

두 번째는 명사로만 이루어진 명사절과는 달리 동사와 명사로 이루어진 명사절은 제대로 처리하지 못한다는 것이다. “New York”을 나누지 말고 한 덩어리로 분석해야 하는 것처럼, “riding a bike”, “cutting a cake” 같은 경우도 한 덩어리로 처리되는 것이 의미상 훨씬 적절하다. 하지만 분석기에서는 이러한 단어들의 조합을 하나의 개념으로 처리하지 못하고 각각의 개념들로 분해해서 독립된 이벤트 구조로 처리한다.

세 번째는 대명사에 대한 처리의 부재이다. 자연 언어 텍스트에는 다양한 용법으로 사용된 대명사가 많이 발견된다. 문장의 의미를 정확히 처리하기 위해서는 이러한 대명사가 지시하고 있는 올바른 대상을 찾아내어 연결시켜 주어야 한다.

네 번째는 ‘Link Grammar’ 영어 구문 분석기가 문법적으로 정확한 문장만을 처리하도록 설계되었다는 점이다. 문법적 오류를 수정하거나 복구할 수 있는 처리가 구현되지 않았기 때문에 모델에 입력되는 텍스트는 철자나 문법, 띄어쓰기, 구두점이 완벽한 문장이라고 가정된다.

5. 분자컴퓨터를 이용한 정서 인식 모델

정서 인식 모델은 DNA 컴퓨터를 사용하여 주어진 텍스트로부터 정서를 추론해낸다. 모델은 텍스트에서 전처리 과정을 거쳐 추출된 단어집합을 입력으로 받아들여 전체 의미 네트워크에서 관련된 단어들을 걸러내게 된다. 걸러진 결과는 실제 세상의 일반지식을 반영하도록 확률적으로 표현된 단어들 사이의 연관 강도에 좌우되는데, 각각의 단어에 지정된 정서 값이 이러한 확률적 관계에 의해 반영되는 정도로부터 입력에 대한 출력 정서 값을 계산할 수 있다.

5.1 DNA 컴퓨터의 구성

DNA 분자 컴퓨터는 단어를 나타내는 가닥과 단어 사이를 연결하는 가닥의 2가지 단일 DNA 가닥(single-strand)으로 구성된다. 두 가지 단일 DNA 가닥들은 서로 연결될 수 있도록 상보적인 염기서열을 가지도록 설계되어서 시험관의 온도가 적당해지면 염기들이 서로 수소 결합을 함으로써 DNA 분자들로 이루어진 이중 나선 구조가 생성된다. 열역학 법칙에 따라 시험관의 온도가 높으면 분자의 운동에너지가 수소 결합보다 강해져서 단일 DNA 가닥들로 분리되려는 경향이 생기고, 반대로 시험관의 온도가 낮아지면 운동에너지보다 수소 결합의 결합력이 강해져서 이중 나선 DNA 가닥들로 존재하려는 경향이 강해진다.

결과 이중 나선 DNA 가닥들의 분포는 각 단일 DNA 가닥들의 화학적 친화도와 분포 개수에 의존적이다. 화학적 친화도는 분자의 구조나 조성을 변화시켜야 하기 때문에 실질적으로 사용할 수 없는 고정된 값이기 때문에 여기서는 분자들의 분포를 조절하는데, 지식은 분자들의 분포에 저장되므로 자연스럽게 확률적인 형태로 표상되고 통계물리학적 특징을 가지게 된다. 이러한 시험관내의 분자들의 확률 분포를 학습과 추론에 이용한 것이 PLM 알고리즘이다.

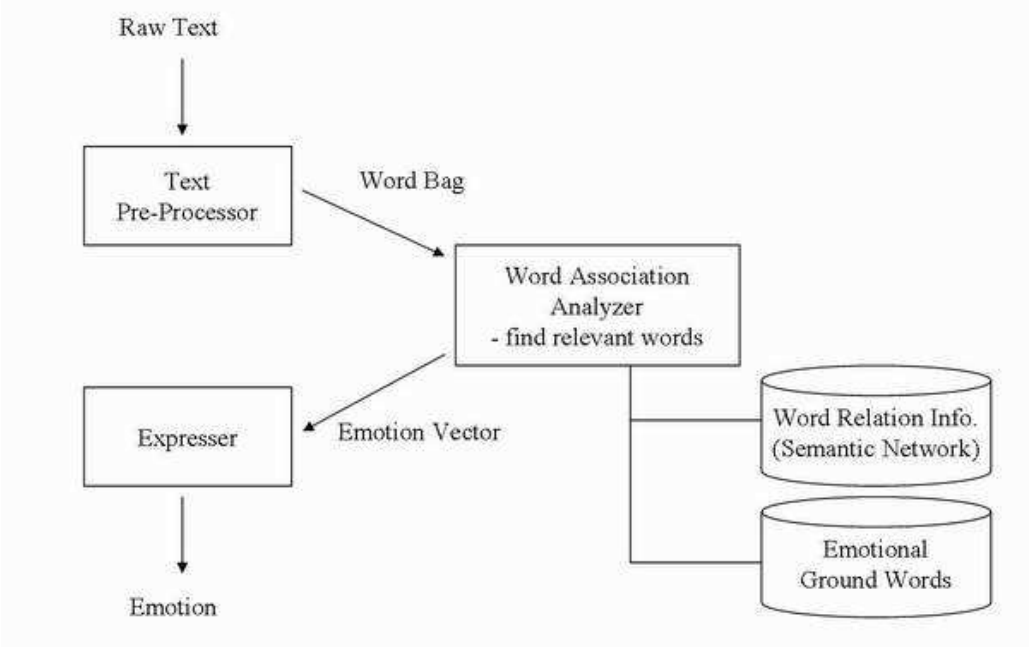

그림 5 Model Architecture

5.1.1 단어 DNA 가닥

단어를 나타내는 DNA 가닥으로 각각의 염기서열의 형태에 따라 대응되는 단어가 달라진다. 따라서 모델에서 처리할 수 있는 단어의 개수와 같은 수의 염기서열의 종류가 존재하게 된다. 하나의 단어 DNA 가닥에는 단어를 나타내는 염기서열 정보에 정서를 나타내는 정보가 덧붙여져서 인코딩되는데 단어의 정서 값은 덧붙여진 정서 값 태그(emotion value tag)의 분포에 의해 확률적으로 표현된다.

정서 값 태그에는 모두 7가지 종류가 있다. Happy, Sad, Angry, Fear, Disgust, Surprising의 6가지 기본 정서를 나타내는 값과 정서가 없는 경우의 Zero를 더해서 총 7가지인데, 각각의 단어 DNA 가닥은 이 중에서 하나의 태그를 가지고 인코딩된다. 단어가 가지는 정서 값은 무작위로 한 개의 단어 DNA 가닥을 뽑았을 때 그 가닥이 특정 정서를 나타내는 태그일 확률과 같다. 예를 들면 800개의 Happy 태그를 가진 가닥들과 4200개의 다른 태그를 가진 가닥들이 있다면, Happy 태그를 뽑을 확률은 0.16이고 이것이 Happy에 대한 정서 값이 된다.

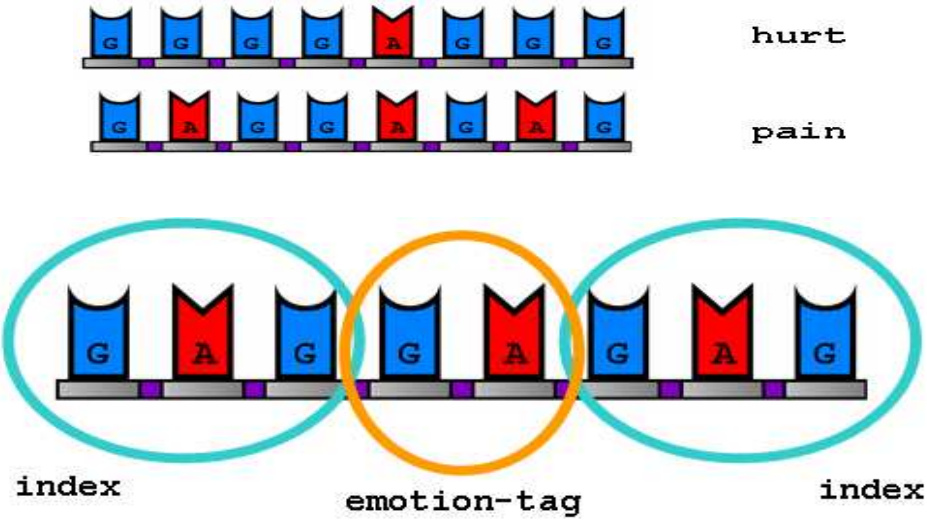

그림 6 단어 DNA 가닥

단어의 정서 값을 벡터로 표현하면 아래와 같은 형태가 된다.

$$v(Concept) = [P(Happy), P(Sad), P(Angry), P(Fear), P(Disgust), P(Surprising)]$$

벡터의 각 원소들은 다음 식을 이용하여 계산할 수 있다.

$$P(Emotion) = \frac{n(Emotion)}{n(\sim Emotion) + n(Emotion)} = \frac{n(Emotion)}{n(Total)}$$

단어 DNA 가닥에 덧붙여진 정서 값 태그의 분포를 조절함으로써 단어의 가지는 정서 값을 확률적으로 표현할 수 있음을 보였는데, 각 단어 별로 분포를 어떻게 결정할 것인가도 대단히 중요한 문제이다. 여기에서는 감정

적 어휘(affective lexicon)에 대한 심리학 연구[27]를 참고하여 직접적인 감정적 표현에 대한 개념에만 정서 값을 부여하고 그렇지 않은 개념에는 아무런 정서를 부여하지 않았다.

Happy: [1.0, 0, 0, 0, 0, 0]

Surprising: [0, 0, 0, 0, 0, 1.0]

Cancer: [0, 0, 0, 0, 0, 0]

Party: [0, 0, 0, 0, 0, 0]

그리고 이중 나선 구조를 만드는 과정에 단어 DNA 가닥의 개수가 영향을 주지 않게 하기 위해 각 단어 DNA 가닥들의 개수는 모두 동일하도록 고정하였다. 즉 단어 DNA의 정서 값 분포는 인코딩된 정서 값 태그의 분포에 의해 결정되지만, 한 종류의 단어에 대한 DNA 가닥의 개수는 일정하다.

$$For\ A,B \in \{all\ word\},\quad if\ n(x)\ is\ the\ number\ of\ word\ sequence$$

$$n(A) = n(A \quad with\ Happy) \quad + \ldots + \quad n(A \quad with\ Surprising)$$

$$n(B) = n(B \quad with\ Happy) \quad + \ldots + \quad n(B \quad with\ Surprising)$$

$$n(A) = n(B) \quad always\ !$$

[27] [Johnson-Laird, (1989)] , [Clore, Ortony, & Foss, (1987)], [John, (1988)] 참조.

5.1.2 연결 DNA 가닥

연결 DNA 가닥은 두 개의 단어 DNA 단일 가닥들을 서로 연결해주는 역할을 한다. 연결하려는 단어 DNA 가닥들의 상보적 염기서열을 가지고 있어서 서로 다른 두 개의 단어가 하나의 이중 DNA 가닥으로 결합할 수 있도록 해준다. 염기서열의 종류에 따른 연결 강도가 일정하다고 가정하면 서로 다른 2개의 단어 DNA 가닥이 적절한 연결 DNA 가닥을 통해 결합될 확률은 단어 DNA 가닥들이 존재할 확률과 그 단어 가닥들을 연결해주는 적절한 연결 DNA 가닥이 존재할 확률에 의해 결정된다. 그리고 각각의 단어 가닥들의 개수는 일정하다고 가정한다면 두 단어가 서로 결합될 확률은 단지 연결 DNA 가닥의 확률에 의해 결정된다. 따라서 서로 밀접한 두 단어를 연결해주는 연결 DNA 가닥이 시험관내에 많이 존재하면 서로 밀접한 두 단어가 서로 결합할 확률은 높아질 것이다. 반대로 서로 관계없는 두 단어를 연결해주는 DNA 가닥의 개수를 줄여주면 서로 상관없는 두 단어가 결합할 확률은 낮아질 것이다. 연결 DNA 가닥들의 분포는 각 단어들 사이의 연관 관계를 확률적으로 표현한다.

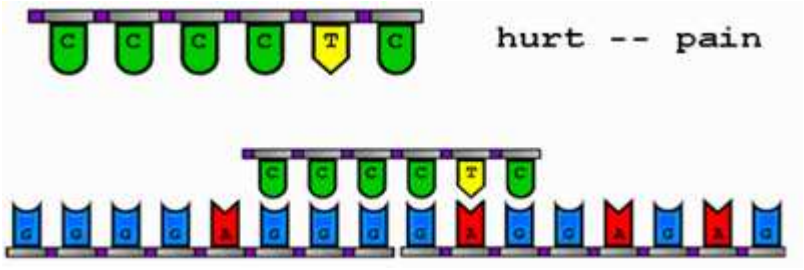

그림 7 연결 DNA 가닥

SPIDER, CRAWL과 FLY라는 3개의 단어 DNA 가닥과 SPIDER-CRAWL, SPIDER-FLY의 2개의 연결 DNA 가닥이 각각 100개씩 존재한다고 생각해보자. 연결 DNA 가닥의 개수가 같아서 SPIDER에 CRAWL이나 FLY 개념이 연결될 확률이 같다면, SPIDER에 CRAWL이 결합한 경우와 SPIDER에 FLY가 결합한 경우는 평균적으로 각각 50개 정도일 것이다. 하지만 SPIDER-FLY의 개수를 10개 정도로 줄이고 SPIDER-CRAWL의 개수를 190개로 늘여주면, SPIDER에 FLY가 결합된 경우는 현저히 줄어들 것이다.

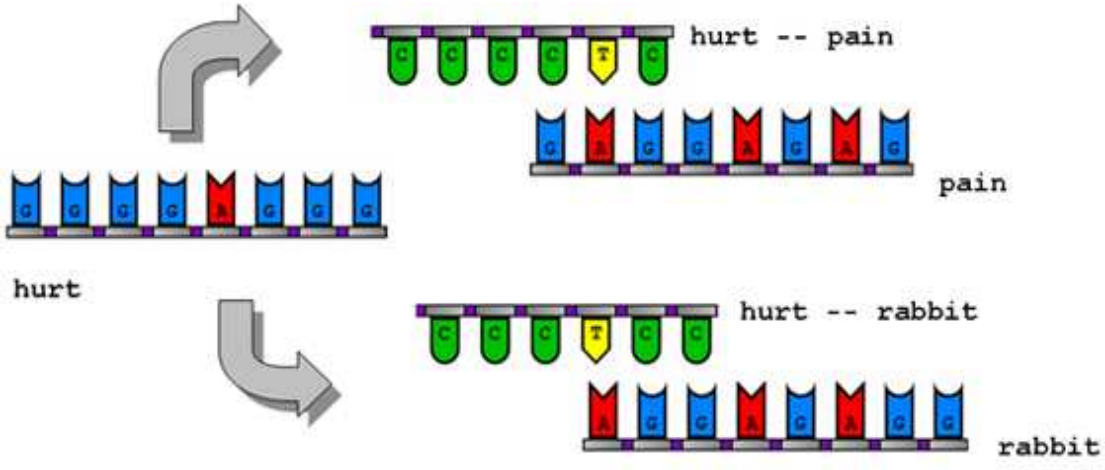

그림 8 단어 DNA가닥과 연결 DNA가닥의 결합

연결 DNA 가닥의 분포는 각 개념들 사이의 연관성을 나타내는데, 이 연관성은 실제 세계의 일반지식(common sense)을 반영하도록 학습되어야 한다. 학습은 하나의 단어주머니(bag of words)가 학습 예제가 되는데, DNA 분자들이 들어있는 시험관에 단어집합에 포함된 단어 인덱스들을 일정한 수를

복제해서 넣어주고 완전한 이중 나선 구조가 생성된 DNA 가닥들만 분리해 내면 이들은 주어진 단어집합에 대해 모든 가능한 양자(bilateral) 관계에 대한 연결 DNA 가닥이 된다. 분리된 연결 DNA 가닥들을 일정한 비율로 복제해서 다시 시험관에 넣어서 분포 확률을 강화시킨다. 이러한 과정을 다른 학습 예제에 대해서도 계속 반복하면 연관성이 있는 개념간의 연결 DNA 가닥들의 농도는 점점 증가하고 그렇지 못한 경우에는 농도가 점차 감소하게 된다.

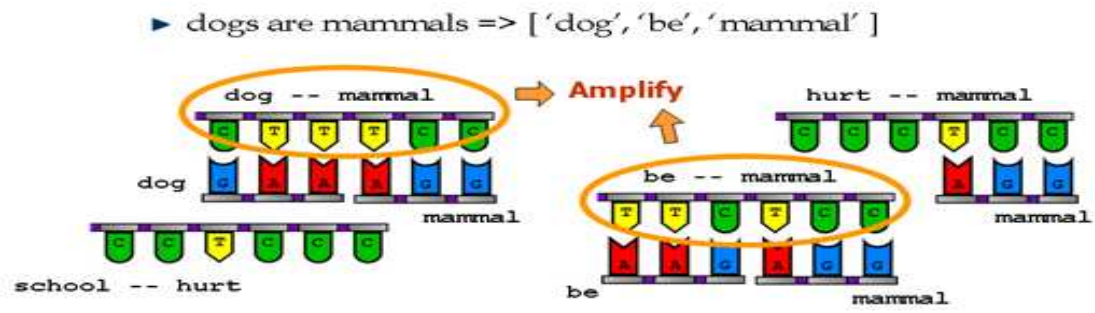

그림 9 연결 DNA 가닥의 확률 분포 조절

학습 알고리즘을 간략하게 표현한 것은 다음과 같다. 라이브러리 L은 연결 DNA 가닥의 전체 집합을 나타내고 X와 Y는 연결하려는 단어들을 나타낸다.

1: Let the library L represent the current empirical distribution $P(X, Y)$

2: Get a training example (x, y)

3: Update L

- $L_n \leftarrow L_{n-1} + d(u,v)$ for u = x, v = y for $(u,v) \in L_{n-1}$

4: Goto step 2 if not terminated

** d(u,v) denotes the number of copies of (u,v)

이 과정에서 d(u,v)는 검출된 연결 시퀀스를 얼마나 복제해서 넣어줄 것인지에 관한 학습율(learning rate)과 관련이 있다. 이 값이 작으면 천천히 배우지만 수렴할 가능성이 높고 값이 크면 하나의 학습 예제에서 빨리 배우지만 수렴할 가능성은 낮아진다. 이 값은 학습 알고리즘의 안정성(stability)와 깊은 관련이 있다.

5.2 DNA 컴퓨터를 통한 계산 과정

5.2.1 결과값 추론

단어 DNA 가닥들과 연결 DNA 가닥들을 같은 시험관에 넣어주고 온도를 천천히 낮춰주면(annealing) 단어 가닥들과 연결 가닥들은 이중 나선 구조를 이루면서 연쇄적으로 결합된다. 이렇게 결합된 DNA 이중 가닥의 길이

는 시험관내의 온도에 따라 결정되는데 섭씨 95도의 시험관에서는 결합이 이루어지지 않고 모든 분자가 단일 가닥의 형태로 존재하지만, 온도가 내려갈수록 이중 가닥을 형성할 확률이 높아져 섭씨 5도 가량에서는 결합 가능한 모든 분자들이 결합하여 기다란 하나의 DNA 이중 가닥을 형성한다.

시험관에서의 계산은 시험관을 적절한 온도로 조절하여 n개의 가닥들이 이어지는 이중 나선 구조가 생성되게 한 후 입력된 단어집합에 있는 단어를 하나라도 포함하는 모든 DNA 가닥들을 분리해내는 과정과 분리된 DNA 가닥들에서 정서 값 태그들의 개수를 세는 과정으로 이루어진다. 첫 번째 과정은 일반지식을 나타내는 전체 DNA 집합에서 특정 단어와 연관된 단어들로 구성된 부분 DNA 집합을 생성하는 것과 같다. 그리고 두 번째 과정은 입력 단어집합으로부터 생성된 부분 집합에서 특정 정서 값 태그를 가지는 DNA 가닥을 magnetic bead를 이용하여 분리함으로써 간단하게 처리될 수 있다.

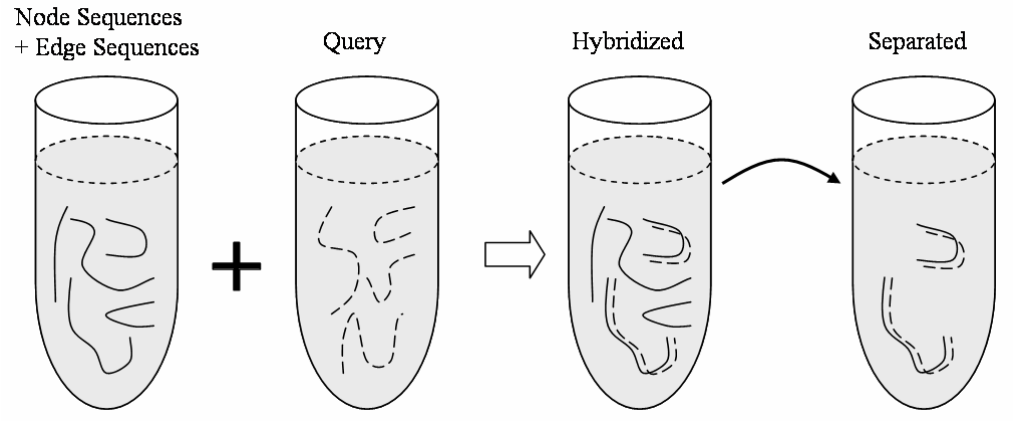

그림 10 시험관에서의 계산 과정

전체 DNA 집합에서 단어 DNA에 인코딩된 정서 값 태그의 개수 분포가 [1000, 1000, 1000, 1000, 1000, 1000]이라고 가정해보자. 이것의 의미는 전체 DNA 가닥에서 Happy, Sad, Angry, Fear, Disgust, Surprising 각각에 대해 1000개씩의 분자가 존재한다는 것이다. 입력된 단어집합에 의해 생성된 부분집합에서 정서 값 태그의 분포가 [300, 120, 500, 300, 100, 100]이었다면, 이 입력에 대한 출력 정서 벡터는 [0.3, 0.12, 0.5, 0.3, 0.1, 0.1]이 된다. 이 벡터는 부분집합으로 분리하는 과정에서 얼마나 많은 정서 값 태그들이 전체 집합에서 걸러졌는지에 대한 확률도 되는데, 출력 정서 분포는 직접적인 감정적 어휘와 관련된 단어 DNA 가닥과 연결된 빈도에 상관이 있고 이러한 빈도는 단어와 단어 사이를 이어주는 연결 DNA 가닥의 분포에 상관이 있다. 주어진 입력에 대해 출력 정서 벡터를 계산하는 추론 과정은 조건부확률을 이용하여 나타낼 수 있는데 이를 수식으로 표현하면 다음과 같다.

입력 패턴을 X라 하고 출력 벡터를 Y라 하면

$$Y = [y(Happy), y(Sad), y(Angry), y(Fear), y(Disgust), y(Surprising)]$$

$u \in \{Happy, Sad, Angry, Fear, Disgust, Surprising\}$ 인 u 에 대해

$$y(u) = P(X \mid u)$$

이고,

베이즈 정리에서 $P(u, X) = P(u \mid X)P(X) = P(X \mid u)P(u)$ 이므로,

$$y(u) = P(X \mid u) = \frac{P(u \mid X)P(X)}{P(u)}$$

가 된다.

전체 DNA 분자 라이브러리를 L 이라 하고

패턴 X에 대해서 분리된 DNA 분자들의 부분 라이브러리를 M 이라 하면,

완전반응(perfect-match) 조건하에서,

$P(X)$ 는 L 에서 X의 확률이고 L^X 가 L 에서 X인 분자들을 나타낼 때,

$$P(X) = \frac{|L^X|}{|L|} = \frac{|M|}{|L|}$$

$P(u)$ 는 L 에서 u 를 가지는 분자의 확률이고, 아래 식에서 $|L^u|$ 는 이미 알고 있는 값이다.

$$P(u) = \frac{|L^u|}{|L|}$$

$P(u \mid X)$ 는 M 에서 u 를 가지는 분자의 확률로

$$P(u \mid X) = \frac{|M^u|}{|M|}$$

가 되는데, $|M^u|$ 는 M 에서 u 인 분자들의 개수로 bead separation을 통해 쉽게 알 수 있는 값이다.

조건부확률 식에 대입해서 정리하면 다음과 같다.

$$P(X|u) = \frac{P(u|X)\,P(X)}{P(u)} = \frac{|M^u| / |M| \cdot |M| / |L|}{|L^u| / |L|} = \frac{|M^u|}{|L^u|}$$

따라서 전체 라이브러리의 정서 분포와 입력 X에 대한 부분 라이브러리의 정서 분포의 비율은 $P(X|u)$를 통해 계산될 수 있다.

5.2.2 의미 네트워크(semantic network)

단어들 사이의 관계가 분자들의 개수에 의해 확률적으로 표현되어있는 DNA 집합은 앞서 언급하였듯이 사실 의미 네트워크 그 자체이다. 단어 DNA와 연결 DNA 분자들의 분포를 완전히 파악할 수 있다면 이로부터 네트워크 구조를 유도해낼 수 있다.[28] 네트워크 구조가 단어들 사이의 연관 관계를 나타낸다면, 주어진 단어의 정서적 의미를 계산하는데 어떤 단어의 정서 태그 분포가 얼마나 반영될 것인가는 단어간의 연관의 강도에 비례한다.

[28] [Zhang, B. T. & Jang, H. Y., (2004)] 참조.

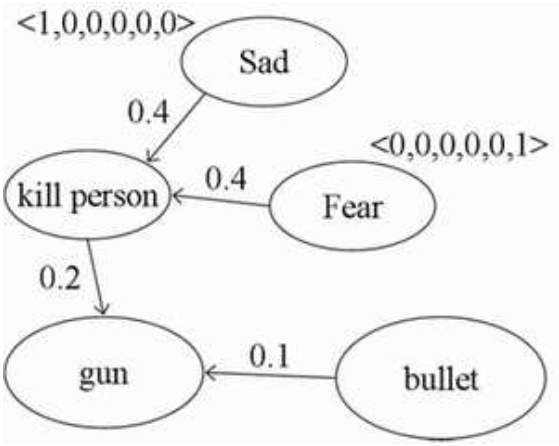

그림 11 Semantic network에 의한 분석의 간단한 예시

위 그림에서만 본다면 kill person은 0.4의 확률로 sad와 fear와 결합하므로, 만약 1000개의 kill person 단어 가닥이 있다면 확률적으로 각각 400개씩의 sad와 fear 가닥과 결합할 것이다. 따라서 kill person의 정서 태그 분포는 <0.4, 0, 0, 0, 0, 0.4>가 된다. Gun 단어 가닥은 0.2의 확률로 kill person과 결합하므로 1000개의 gun 단어 가닥이 포함된 DNA 이중 가닥을 분리하였다면 그 중 kill person이 포함된 이중 가닥은 200개이고 그 중 sad와 fear 가닥이 결합된 경우는 각각 80개가 될 것이다. 따라서 gun의 정서적 의미의 계산 과정에서 kill person의 태그 분포는 0.2 정도가 반영되고 bullet은 0.1, fear와 sad는 각각 0.08의 정도로 기여를 하게 될 것이다.

6. 시뮬레이션 및 결과

6.1 시뮬레이션 데이터 및 절차

시뮬레이션에서는 먼저 Open Mind Common Sense Raw Text DB[29]에서 약 $5*10^5$개의 문장에 대해 전처리 과정을 거쳐 단어들을 추출하였다. 추출된 단어들에서 가장 빈도수가 높은 5000개의 단어를 선택하여 5000가지 종류의 단어 DNA 가닥을 만들었고, Common Sense DB의 문장들을 이용해 5000개의 단어에 대해 그 연관 관계를 학습하였다. 정서적 단어는 약 40개가 포함되었고 각각 표현에 적당한 정서 값 태그 분포를 할당하였다. 단어 가닥들은 종류별로 10^5개, 연결 가닥들은 종류별로 10^4개로 초기화하였고, 연결 가닥들은 한 번씩 강화가 될 때마다 현재 크기의 N%씩 개수를 증폭시켜 시험관에 다시 넣어주었다.

실제 시뮬레이션은 분자컴퓨터가 아닌 일반적인 컴퓨터 프로그램을 이용하여 수행되었고, 구체적인 시뮬레이션 절차는 다음과 같다.

1) 학습 데이터를 전처리하고 빈도수에 따라 모델에서 사용할 5000개의 단어를 선정한다.
2) 단어 DNA 분자들의 분포를 결정한다. (Affective Lexicon 연구에서)
3) PLM 알고리즘으로 연결 DNA 분자들의 확률 분포를 학습한다.

---

[29] [Singh, P. (2002)]와 http://commonsense.media.mit.edu 참조.

4) 테스트 데이터에 대해 계산된 정서를 실제 사람에게 제시하여 적절한 정서인지 아닌지 판단하게 하였고, 그 결과를 기준으로 모델의 정확도를 판단하였다.

6.2 시뮬레이션 결과

6.2.1 단어 DNA 분포

부록. 1 참조.[30]

6.2.2 연결 DNA 분포

입력된 조건들에 대해 가능한 모든 양자간(bilateral) 연결에 대해 0.5%의 비율로 개수를 증가시켰다. 다음 페이지의 그래프는 의미 네트워크의 노드들에 대해서 노드의 차수에 대한 노드의 개수를 나타낸 것이다. X축은 노드의 차수(degree)로 노드에 연결되는 다른 노드들의 총합을 나타내고, Y축은 특정 차수를 가지는 노드의 분포이다.

[30] [Johnson-Laird, (1989)], [John, (1988)] 참조.

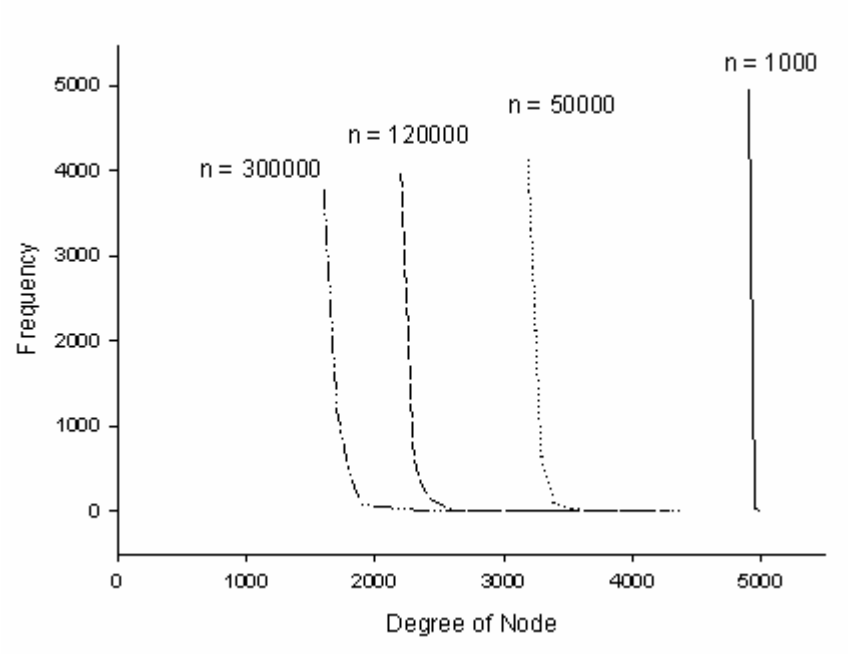

그림 12 학습 예제 크기 별 노드 차수 분포

초기에는 모든 노드들이 같은 수의 연결을 가지기 때문에 하나의 X값에서 전체 네트워크의 노드 개수만큼의 Y값을 가지는 델타 함수 형태의 그래프로 그려진다. 이러한 네트워크에서 PLM에 의한 학습은 가지치기 과정과 같은데, 강화가 되지 못한 노드간의 연결은 상대적으로 점차 약화되어 끊어지게 된다. 차수에 대한 노드의 분포는 파워 분포로 나타나는데 학습이 진행될수록 노드들의 평균 차수는 계속 감소하게 된다. 과다하게 학습이 일어나면 학습 예제에서 빈도가 높은 경우를 제외하고 전부 연결이 끊어지게 되는데, 서로 관련이 없는 노드들 뿐만 아니라 학습 예제에서 나타났지만 빈도가 낮은 관계도 여기에 포함되게 된다. 여기서는 경험적으로 노드의 평균 차수가 1500정도가 될 때까지 학습을 진행하였다.

```
## Relevant Concept List For
   50:money

31:  buy        0.00229
12:  make       0.00094
16:  get        0.00086
71:  pay        0.00079
452: steal      0.00076
228: spend      0.00064
72:  need       0.00061
129: bill       0.00060
756: jail       0.00054
439: earn       0.00053
46:  work       0.00051
365: save       0.00050
253: job        0.00050
54:  give       0.00048
557: business   0.00047
24:  take       0.00045
1458:wallet     0.00045
620: poker      0.00044
173: win        0.00043
227: cookie     0.00043
794: cost       0.00043
1357:bank       0.00043
20:  play       0.00043
```

```
## Relevant Concept List For
   85:game

20:  play       0.01128
380: chess      0.00067
173: win        0.00059
31:  buy        0.00057
1902:with your friend   0.00048
12:  make       0.00046
122: lose       0.00046
152: fun        0.00044
19:  true       0.00044
66:  learn      0.00042
761: rule       0.00041
620: poker      0.00041
308: score      0.00041
43:  watch      0.00040
651: homer      0.00040
224: basketball 0.00040
1001:skill      0.00039
1124:golf       0.00039
471: hockey     0.00039
167: enjoy      0.00039
50:  money      0.00039
106: move       0.00039
64:  dog        0.00039
```

그림 13 의미 네트워크에서 관련 단어 추출의 예

이렇게 학습된 연결 DNA 분포는 임의의 노드에 대해 연결된 다른 노드들과 연결된 강도를 저장하고 있다. 그림 13은 질의 단어에 대한 의미 네트워크의 출력의 몇 가지 예이다.

6.2.3 출력 정서 벡터

주어진 입력에 대해 모델이 산출한 정서 벡터의 정확도는 결과에 대한 사람의 평가에 의해 측정되었다. 모두 5명이 모델의 산출 결과를 평가하였

고 정확도는 전체 사례 중에서 긍정적(YES) 반응을 보인 사례의 비율로 하였다.

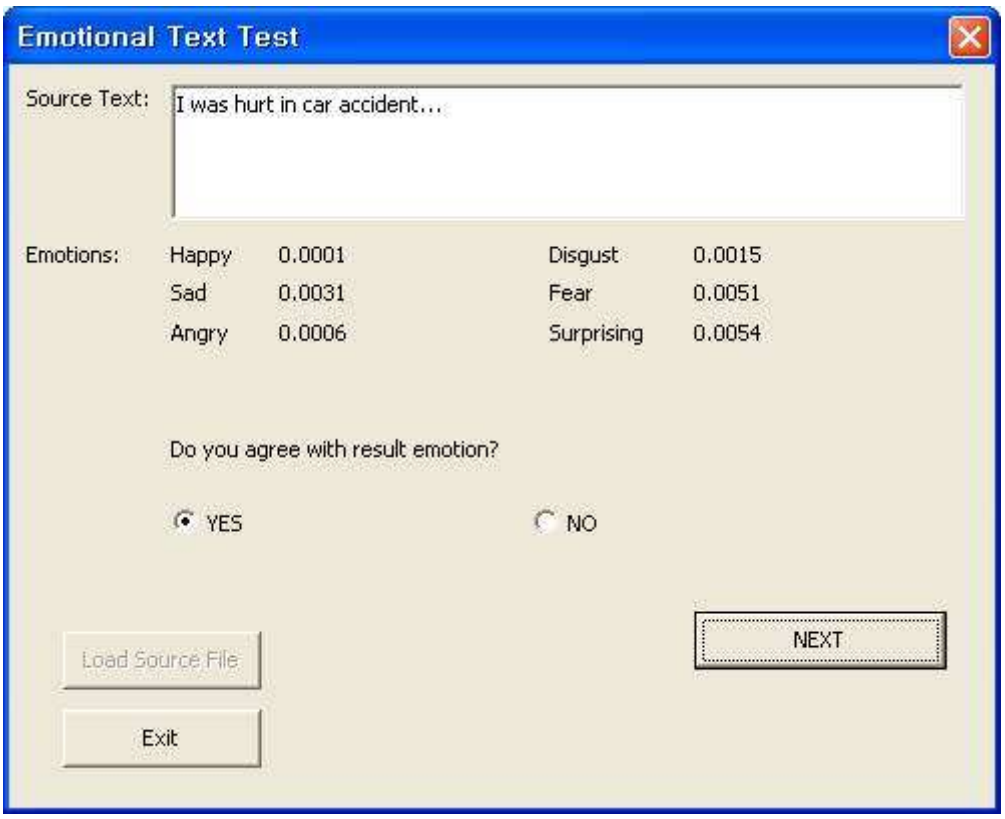

그림 14 모델의 산출 결과 평가 프로그램

모델의 테스트는 3가지 텍스트 형태에 대해 이루어졌다. 먼저 단어 하나에 대한 실험은 모델의 전처리 과정에서 선정된 5000개의 단어 중에서 임의의 단어들을 선정하여 사용하였다. 문장 하나에 대한 실험은 4장에서 소

개한 분석과정을 거쳐 추출된 단어 집합을 사용하였고 주술구조가 하나만 존재하는 단순한 문장 만을 사용하였다. 문단에 대해서는 출현 빈도가 높은 순으로 모델에서 처리할 수 있는 5000개에 포함되는 단어들을 최대 4개까지 선정하여 단어 집합을 생성하여 모델에 사용하였다.

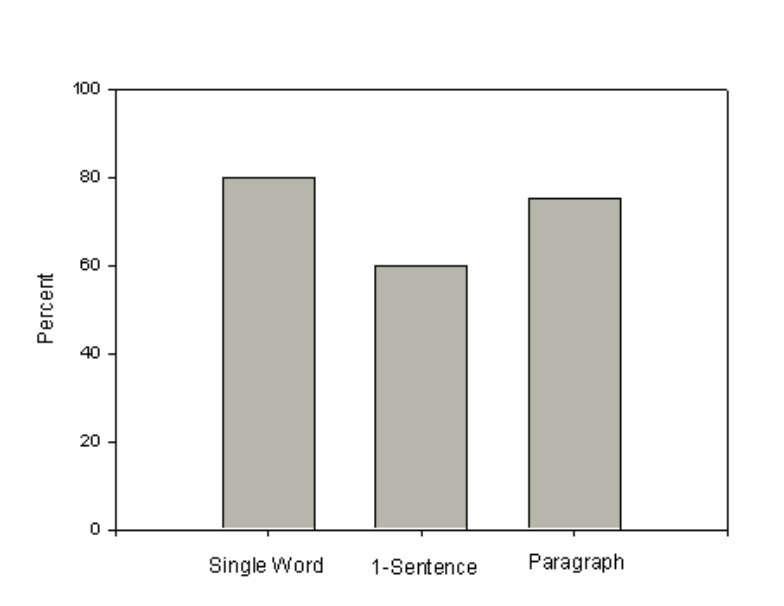

그림 15 텍스트 형태 별 정확도

단어 하나에 대한 결과는 비교적 정확도가 높지만, 문장에 대한 결과는 정확도가 좋지 않다. 문장에서 추출된 개념들에 대한 의미적 정보가 단순하게 집합을 이용한 표현에서는 거의 소실되기 때문이다. 집합내의 개념들이 거의 비슷한 정서적 의미를 나타내는 경우에는 이러한 접근 방법이 문제가 없지만, 서로 다른 정서적 의미를 나타내는 개념들이 뒤섞여 있는 경우나

개념들의 조합에 의해 전혀 다른 의미를 생성하는 경우에는 이러한 방법으로는 추론할 수 없다.

‘철수가 웃는다.’

‘철수가 혼자서 웃는다.’

위 간단한 예에서 볼 수 있듯이 단지 ‘혼자서’ 한 단어가 추가되었음에도 불구하고 정서적 의미는 완전히 달라질 수 있다.

보다 긴 글에 대해서는 주제어들을 추출하는 과정이 통계적으로 이루어지므로 문장에서처럼 단어가 한 번 나타났다고 의미가 급격하게 바뀌는 경우가 드물다. 이처럼 단어의 빈도를 이용해 주제어를 추출할 경우 텍스트의 길이가 길어질수록 추출된 주제어는 텍스트의 의미와 밀접한 관련을 가질 확률이 높아지지만, 통계적 의미는 문장을 처리해서 얻을 수 있는 날카로운 의미와 달리 평균적이고 완만한 의미를 가지는 한계가 있다.

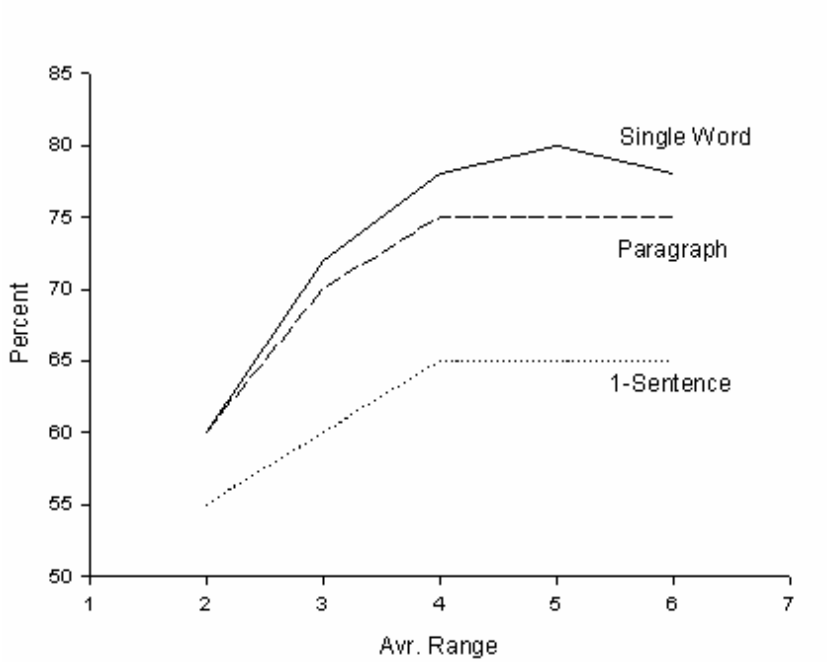

그림 16 검색 범위 별 정확도

검색 범위는 생성된 분자 구조의 길이와 관련이 깊다. 분자 구조의 길이가 길어지면 입력된 단어와 바로 연관되는 단어만 사용하는 것이 아니라 여러 단계를 걸쳐 연관되는 단어들도 결과 값 계산에 반영하게 된다. DNA 컴퓨터에서는 분자 구조의 길이를 시험관의 온도를 이용하여 조절할 수 있다. 온도를 낮출수록 분자 구조가 평균적으로 길어져서 결과적으로 넓은 검색 범위를 갖게 된다. 검색 범위가 넓어질수록 모델은 입력 단어집합의 일반화된 의미에 대해서 추론을 하게 되므로 더 안정적으로 동작할 수 있다. 그러나 출력 정서 분포 벡터에서 각 정서 별 차이가 줄어들어서 완만한 값을 가지게 되고, 지나치게 넓은 검색 범위는 모델이 입력 텍스트에서 거의 모든 종류의 정서를 인식하게 할 것이다.

6.2.4 의미 네트워크

전체 분자 집합에서 입력 조건에 맞는 분자들을 걸러낸 부분 집합도 서로 네트워크 구조를 가짐은 앞서 언급하였다. 걸러진 분자 구조들로부터 의미 네트워크가 그려지는 과정은 다음과 같다. 먼저 DNA 이중 나선 구조의 한쪽 끝에서부터 단어를 차례대로 읽어 들여서 네트워크상의 단어들과 대응시킨다. 네트워크에 존재하지 않는 경우에는 새로 단어 노드를 그려주고 이미 존재하는 단어이면 출현 빈도수를 하나 높여준다. 네트워크상에서 단어들을 순서대로 지나가면서 없는 링크는 새로 만들어주고, 이미 존재하는 링크에는 빈도수를 1씩 더해준다.

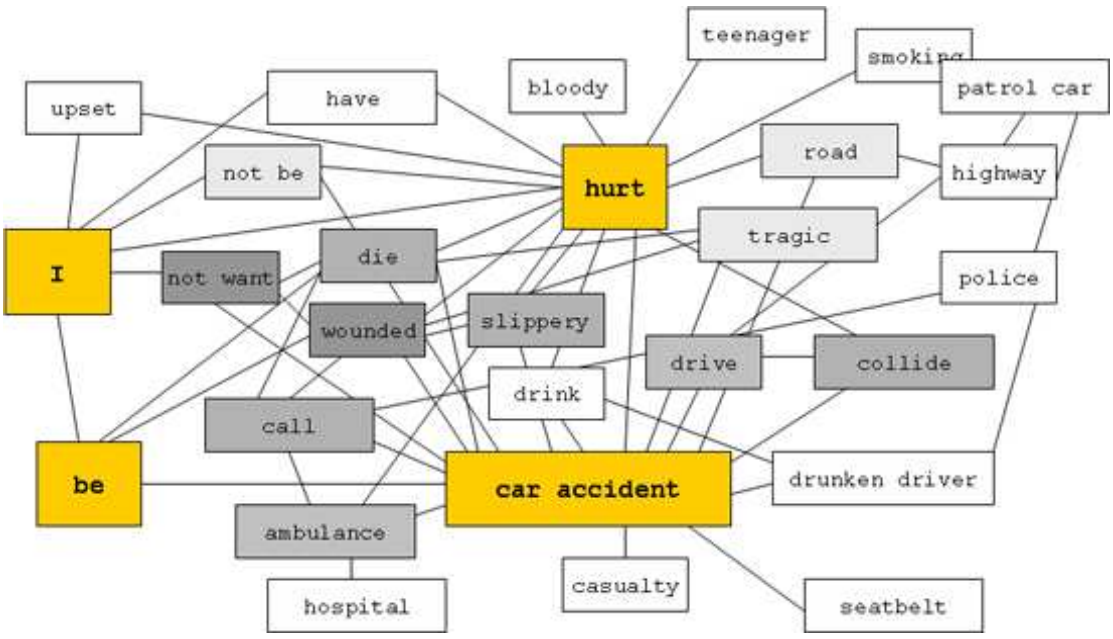

그림 17 'I was hurt in car accident'의 결과 의미 네트워크

위 네트워크는 [‘I’, ‘be’, ‘hurt’, ‘car accident’]에 대한 결과 분자 집합을 가지고 생성한 결과이다. 노드가 많아서 네트워크가 매우 복잡하므로 일정 빈도수 이상의 노드와 링크 만을 나타내었는데, 노란색 노드는 입력된 ‘I’, ‘be’, ‘hurt’, ‘car accident’을 나타내고 회색 노드들은 추론에 사용된 관련 개념들을 나타내는데 색이 진할수록 빈도수가 높은 개념이고 결과값에 높은 영향력을 가진다.

7. 결론

7.1 요약

지금까지 DNA 컴퓨터를 이용하여 의미 네트워크를 구성하고 텍스트의 정서를 추론하는 모델을 설명하였다. 먼저 의미적인 관련성이 정서를 추론하는데 어떻게 이용될 수 있는지 심리학적 선행 연구를 통해 살펴보았고, 사람의 자연 언어를 대상으로 하는 기존의 인공지능 분야의 연구들을 살펴보고 사람의 인지 과정 특히 정서의 문제를 모사하려는 비튜링기계로써의 DNA 컴퓨터가 기존의 방법론에 대해 가지는 강점을 살펴보았다. 본 모델은 단어에 대한 정보와 단어 사이의 의미적 관계에 대한 정보를 이용해 간단한 전처리 과정을 거친 텍스트에 대해서 정서적 의미를 추론해낸다. 분자 컴퓨터를 이용한 계산 과정은 개념을 나타내는 DNA 분자들이 연쇄적으로 자기조립되어 결과 구조물을 생성해내는 과정으로 최종 모델의 출력은 PLM 알고리즘에서 분자들의 통계적 특성을 통해 계산되게 된다. 미시적인 분자 사이의 계산 과정은 일차원적인 논리 추론에 가깝지만, 거시적인 모델의 계산 과정은 개념을 나타내는 분자들을 이용해서 의미 네트워크를 생성하는 과정과 같다. 의미 네트워크에서 각 노드들의 관계는 일반적인 지식을 담고 있는 언어 데이터로부터 학습하였고 다른 언어 네트워크에서처럼 자연 발생적인 네트워크가 가지는 scale free 특성을 관찰할 수 있었다. 모델의 성능은 같은 텍스트에 대한 사람의 추론 결과와 비교하여 측정하였는데, 단일 단어

텍스트와 비교적 길이가 긴 텍스트에 대해서는 비교적 좋은 결과를 보였으나 단일 문장 텍스트에 대해서는 추론의 정확도가 떨어졌다.

7.2 연구 의의

인공지능 알고리즘을 이용하여 텍스트에서 정서를 인식하려는 연구는 이미 다양하게 시도되어왔다. 본 모델은 단어의 의미를 확률적으로 표현한다는 점에서 앞서 언급했던 기존 연구에서 통계적인 언어 처리 방법과 매우 유사하다. 하지만 일반적으로 텍스트를 단어 주머니 형태로 처리하는 모델에서 단어 사이의 유사도가 단어 출현 빈도 벡터들의 공간적 거리로부터 계산되는데 비해, 본 모델에서는 단어 사이의 관계에 대한 의미 네트워크를 사용하므로 단어 사이의 공간적 거리와 의미적 거리가 비례한다는 가정으로부터 자유로울 수 있다.

계산 과정에서 사용된 의미 네트워크는 사람의 언어 데이터를 기초로 만들어진 일반적인 지식 구조이다. 일반적인 기계 학습 알고리즘에서 정서를 추론하기 위해서는 주어진 텍스트에 대해서 감독학습을 해야 한다. 따라서 정서를 추론하기 위해 학습된 특화된 모델을 텍스트를 이용하는 다른 문제에 적용시키기는 거의 불가능하다. 하지만 일반적인 지식 구조를 활용하는 본 모델의 추론 방식은 학습된 결과물을 의미 네트워크를 이용하는 다른 비슷한 문제에 대해서 그대로 재사용될 수 있다. 이것은 보다 지능적인 시

스템이 갖추어야 할 특징인 일반지식(common-sense) 기반 문제 해결에 대한 좋은 시도가 될 수 있다.

두 번째 의의는 지식을 저장하고 검색하는 과정을 기존의 튜링 컴퓨터가 아닌 분자 컴퓨터의 확률적 계산 과정을 이용해 구현했다는 점이다. 기존의 전통적인 컴퓨터가 안정적이고 주어진 입력에 대해 안정적인 내부 상태를 가지는데 비해 분자 컴퓨터는 내적 상태가 비선형적이고 계산 과정도 순차적인 튜링 알고리즘으로 설명될 수 없는 동역학적 특성을 가진다. 본 모델은 이러한 분자컴퓨터의 동적인 특성을 잘 조절하여 수렴된 결과값을 도출해내는 지식의 저장 및 검색 알고리즘으로써 의의가 있다.

세 번째는 단순한 계산 과정들의 조합을 통해 상위 수준의 복잡한 과정을 창발적으로 설명을 시도했다는 점이다. 분자들 사이의 일대일 결합은 단순한 논리의 연속에 불과하지만 수많은 분자들의 분포를 이용함으로써 결과물을 확률적으로 분석할 수 있었다. 다양한 분자들 사이의 의존도 또한 분포를 이용해 확률적으로 표현되기 때문에 실제 분자들은 서로 친화도가 높은 것끼리 연결되려는 작용만 하지만 시험관내에서 전체적인 관점에서 보자면 분자들 사이의 확률 기반 의미 네트워크가 생성되고 이것을 통해 결과값이 계산되는 것처럼 보이게 된다. 그리고 이러한 과정이 외부의 인위적인 규칙이나 예외를 전혀 사용하지 않고 순수하게 상향식(bottom-up)으로 이루어졌다는 점에서 중요한 의미를 가진다.

마지막으로 네 번째 의의는 정서에 대한 본 모델의 접근 방식이 생물학적인 정서 시스템과 유사하다는 점이다. 생물체는 자극에 대해서 복잡한 신

호 처리 과정을 거쳐 특정 정서에 관련된 신호 단백질을 생산하게 되고, 이들은 생화학적 경로(pathway)를 통해 수용기를 가진 세포들에 전달된다. 이러한 단백질에 의한 전달 과정에서 가장 중요하게 작용하는 것은 얼마나 많은 수의 세포들이 신호 단백질의 생성과 수용에 관여하느냐이다. 어느 신경적 경로의 어떤 세포에 의해 생성된 신호 단백질인지에 상관없이 몸의 전체적인 정서 상태는 혈관 속의 신호 단백질의 개수에 의해 좌우된다. 특정 조건의 만족 여부에 따라 정서 상태가 알고리즘적으로 변화하는 것이 아니라 정서 태그를 가지는 노드들에 자극으로 주어진 노드들이 완전히 병렬적으로 연결되는 과정을 통해 전체 태그의 분포가 결정되는 본 모델의 처리 과정은 두뇌에서 신호 단백질을 분비하는 세포들에 자극을 받은 신경 세포들이 병렬적으로 연결되어 혈관 내 단백질 농도가 결정되는 과정과 비교할 수 있을 것이다.

7.3 모델의 한계점

문장을 단어집합으로 표현하는 데에 본 모델의 첫 번째 한계점이 있다. 단어 집합을 이용한 추론 과정은 의미 네트워크에서 각 단어들 사이의 상대적 거리에 의해 이루어지게 되는데, 의미 네트워크에서의 단어들 사이의 거리는 입력된 단어집합의 전체적인 관점에서 볼 때 의미론적으로 강한 의미를 가지지 못한다. 시뮬레이션 결과에서 나타나듯이 단어 하나의 처리나 긴 텍스트에서 통계적으로 추출된 주제어 집합에 대해서는 비교적 높은 정확도

를 보여주는데, 문장의 의미는 개념들 사이의 순서에 큰 영향을 받으며 개념들의 조합에 의한 의미가 따로따로 나누어지면서 소실되는 경우가 있기 때문에 쉽게 오류에 빠진다. 의미 네트워크에서 이러한 단조로운 추론을 극복하기 위한 방안 중의 하나는 술어(predicate)를 사용하여 단어 사이의 관계를 규정하는 것이다.

두 번째는 인식되는 정서가 몇 개의 정서적 표현에 국한된 점이다. 슬픔(sad)에 관련된 정서들은 똑 같은 슬픔이란 단어에 의해 지칭되지만 셀 수 없이 다양한 형태를 가진다. 이 모델에서 분류하는 슬픔은 단지 sad, sadness, grief 같은 단어의 전형적인 의미에 바탕을 두고 있을 뿐이다. 사실 독자가 텍스트를 읽으면서 얻을 수 있는 정서는 처리 수준에 따라 훨씬 다양하다. 문장 구조나 추상적인 이야기 구조에서 표현되는 정서도 있을 것이고, 텍스트의 등장 인물들에 감정 이입을 할 수도 있을 것이다. 이러한 수준의 정서들은 단어 수준에서는 추론해내기 거의 불가능한 것들이다.

7.4 향후 연구

본 연구의 향후 과제에 대해 고려해야 할 점은 다음과 같다.

먼저 텍스트의 의미에 대한 더 나은 확률적 표상을 고안해야 한다. 텍스트에서 추출된 단어들이 공간적 혹은 시간적으로 구조를 이루도록 함으로써 모델의 표현력을 높일 수 있으므로 문장 수준의 짧은 텍스트에 대한 정확한 처리가 가능해진다. 그리고 의미가 하나의 단어에 제한되지 않고 단어

들의 구조로부터 파악할 수 있는 여지가 생긴다. 앞서 언급했듯이 술어구조의 사용은 가장 손쉽게 떠올릴 수 있는 방안인데, 이러한 구조가 분자적인 수준에서 자동적으로 조립되고 해석되는 구현상의 미시적 알고리즘과 함께 시스템 관점에서 동역학적인 창발적 속성에 대해서도 함께 고려해야 한다.

두 번째는 시스템의 시간적 변화에 대한 고려이다. 분자 컴퓨터를 이용한 의미네트워크는 공간적으로 매우 동적인 속성을 가질 뿐 아니라 시간적으로도 끊임없이 변화하는 시스템이다. 텍스트 처리 과정도 과거에 처리한 텍스트가 이후의 처리 과정에 영향을 미치는 시간적으로 밀접한 연관을 가지는 처리 과정이다. 따라서 텍스트의 문맥적 의미 추론을 위해서는 시스템의 시간적 변화에 대한 고려가 필수적이다.

참 고 문 헌

김지수, (2005). 분자 컴퓨팅을 이용한 감성 정보 범주화. 석사 학위 논문, 서울대학교 인지과학 협동과정.

이은석, (2003). DNAgram: Anagram 문제 해결에 관한 분자 컴퓨팅 시뮬레이션 연구. 석사 학위 논문, 서울대학교 인지과학 협동과정.

장병탁, (2005). 나노바이오지능 분자 컴퓨터: 컴퓨터 공학과 바이오 공학, 나노기술, 인지뇌과학의 만남, 한국정보과학회지 바이오정보기술 특집호, 2005년 5월.

Adleman, L. (1994). Molecular Computation of solutions To Combinatorial Problem. Science, 266, 1021-1024.

Anderson, J. R. (1990). The Adaptive Character of Thought. Erlbaum, Hillsdale, NJ.

Barabasi, A. L. & Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 509-512.

Barker, K., Robertson, N., (1997). Selective Processing and Fear of Spiders: Use of the Stroop Task to Assess Interference for Spider-related, Movement, and Disgust Information, Cognition and Emotion, 11(3), 331-336.

Berridge, K. C., & Winkielman, P. (2003). What is an unconscious emotion: The case for unconscious 'liking'. Cognition and Emotion, 17, 181-211.

Bestgen, Y. (1994). Can emotional valence in stories be determined from words? Cognition and Emotion, 8, 21-36.

Brooks, R. A. (1999). Cambrian Intelligence. Cambridge, MA: The MIT Press.

Brooks, R. A. (2002). Flesh and Machines. New York, NY: Pantheon Books.

Clore, G. L., Ortony, A., & Foss, M. A. (1987). The psychological foundations of the affective lexicon. Journal of Personality and Social Psychology, 53, 751-766.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. Psychological Review, 82, 407-428.

Damasio, A. (1994). Descartes' Error, Emotion Reason and the Human Brain. New York: Grosset/Putnam Books.

Dyer, M.G. (1987). Emotions and Their Computations: Three Computer Models. Cognition and Emotion, 1(3), 323-347.

Elliott, C. (1992). The Affective Reasoner: A Process Model of Emotions in a Multi-agent System. PhD thesis, Northwestern University, The Institute for the Learning Sciences, Technical Report No. 32.

Elliott, C. (1993). Using the affective reasoner to support social simulations. In Proceedings of the Thirteenth Annual Joint Conference on Artificial Intelligence, pp 194-200, Chambery, France, August 1993. Morgan Kaufmann.

Ekman, P. (1993). Facial expression of emotion. American Psychologist, 48, 384-392.

Gernsbacher, M. A., Goldsmith, H. H., & Robertson, R. R. W. (1992). Do Readers Mentally Represent Characters' Emotional States?. Cognition and Emotion, 6(2), 89-111.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In Proceedings of the Twenty-Fourth Annual Conference of Cognitive Science Society. George Mason University, Fairfax, VA.

Griffiths, T. L., & Steyvers, M. (2003). Prediction and Semantic Association. Advances in Neural Information Processing Systems, 15, 11-18.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. Psychological Review, 108(4), 814-834.

Hermans, D., Baeyens, F., & Eelen, P. (1998). Odours as Affective-processing Context for Word Evaluation: A case of Cross-modal Affective Priming. Cognition and Emotion, 12(4), 601-613.

Houwer, J. D., Hermans, D., Rothermund, K., & Wentura, D. (2002). Affective priming of semantic categorization responses. Cognition and Emotion, 16(5), 643-666.

Jackendoff, R., (1990). Semantic Structures. Cambridge, MA: The MIT Press.

John, C. H. (1988). Emotionally ratings and free-association norms of 240 emotional and non-emotional words. Cognition and Emotion, 2(1), 49-70.

Johnson-Laird, P. N., & Oatley, K. (1989). The Language of Emotions: An Analysis of a Semantic Field. Cognition and Emotion, 3(2), 81-123.

Kunda, Z. (1999). Social Cognition. Cambridge, MA: MIT Press.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Lewis, M., & Haviland, J. M. (Eds.) (2000). Handbook of Emotions. New York: Guilford Press.

Liu, H., Lieberman, H., & Selker, T. (2003). A Model of Textual Affect Sensing using Real-World Knowledge. Proceedings of the 2003 International Conference on Intelligent User Interfaces (pp. 125-132). Miami, Florida.

Liu, Hugo (2004). MontyLingua: An end-to-end natural language processor with common sense. Available at: web.media.mit.edu/~hugo/montylingua.

Lodish, H. et. al. (2000). Molecular Cell Biology(4th ed.). New York: W. H. Freeman.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., Welch, N. (2001). Risk as Feelings, Psychological Bulletin, 127, 267-286.

McCarthy, J., Minsky, M., Sloman, A., Gong, L., et al. (2002). An architecture of diversity for commonsense reasoning, IBM Systems Journal, 41(3), 530-539.

Mohlman, J., Mangels, J., Craske, M. G., (2004). The Spider Phobia Card Sorting Test: An investigation of phobic fear and executive functioning, Cognition and Emotion, 18(7), 939-960.

Minsky, M. (1986). The society of mind. New York: Simon and Schuster.

Minsky, M. (forthcoming). The emotion machine. Pantheon, not yet available in hardcopy. Drafts available: http://web.media.mit.edu/~minsky/.

Nass, C.I., Stener, J.S., and Tanber, E. (1994). Computers are social actors. In Proceedings of CHI ’94, (Boston, MA), pp. 72-78, April 1994.

Niedenthal, P.M., Halberstadt, J.B., (1999). Emotional Response Categorization. Psychological Review, 106(2), 337-361.

Oatley, K., & Johnson-Laird, P. N. (1989). Towards a cognitive theory of emotions. Cognition and Emotion, 3, 125-137.

Oatley, K., & Jenkins, J. M. (1996). Understanding Emotions. Cambridge, MA: Blackwell.

Panksepp, J. (1998). Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford: Oxford University Press.

Picard, R. W. (1997). Affective Computing. Cambridge, MA: The MIT Press.

Picard, R. W., (1998). Human-Computer Coupling. Proceedings of the IEEE, vol.86, No.8, TR469

Picard, R.W. (1998). Toward Agents that Recognize Emotion. Actes Proceedings IMAGINA, March 1998, pp. 153-165, Monaco.

Salovey, P. and Mayer, J. D. (1990). Emotional Intelligence, Imagination, Cognition and Personality, vol. 9, no. 3, pp. 185-211.

Seligman, M. E. P. (1971). Phobia and preparedness. Behaviour Therapy, 2, 307-320.

Singh, P. (2002). The Open Mind Common Sense Project. KurzweilAI.net. Retrieved from http://www.kurzweilai.net/meme/frame.html?main=/articles/art0371.html

Singh, P. (2002). The public acquisition of commonsense knowledge. Proceedings of AAAI Spring Symposium on Acquiring(and Using) Linguistic(and World) Knowledge for Information Access. Palo Alto, CA: AAAI.

Sloman, A. (1998). Categorical inferences is not a tree: The myth of inheritance hierarchies. Cognitive Psychology, 35, 1-33.

Sloman, A. (2001). Beyond Shallow Models of Emotion. Cognitive Processing, 2(1), 178-198.

Sloman, A. (2004). What are emotion theories about, Multidisciplinary workshop on Architectures for modeling Emotion at the AAAI Spring Symposium at Stanford University in Mar 2004.

Sloman, A. (2005). More things than are dreamt of in your biology: Information processing in biologically-inspired robots. Cognitive Systems Research, 6(2), 145-174.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In A. Healy(Ed.), Experimental Cognitive Psychology and its Applications.

Steyvers, M., Tenenbaum, J. B., (2005). The Large-Scale Structures of Semantic Networks: Statistically Analyses and a Model of Semantic Growth. Cognitive Science, 29, 41-78.

Thagard, P. (1996). Mind: Introduction to Cognitive Science. Cambridge, MA: MIT Press.

Thagard, P. (2002). How molecules matter to mental computation. Philosophy of Science, 69, 429-446.

Watts, F. N., KcKenna, F.P., Sharrock, R., & Trezise, L. (1986). Colour naming of phobia-related words. British Journal of Psychology, 77, 97-108.

White, M. (1996). Automatic Affective Appraisal of words, Cognition and Emotion. 10(2), 199-211.

Winfree, E., Liu, F., Wenzler, L., & Seeman, N. C. (1998). Design and self-assembly of two-dimensional DNA crystals. Nature, 394, 539-544.

Winkielman, P., Berridge, K.C. (2004). Unconscious Emotion. Current Directions in Psychological Science, 13, 120-123.

Zhang, B.-T., & Jang, H. –Y. (2004). A Bayesian Algorithm for In Vitro Molecular Evolution of Pattern Classifiers. Proceedings of the Tenth International Meeting on DNA Computing (pp. 294-303). Milano: Springer.

Zhang, B.-T., & Jang, H. –Y. (2005). Molecular Programming: Evolving Genetic Programs in a Test Tube, The Genetic and Evolutionary Computation Conference (GECCO 2005). (to appear)

Zhang, B.-T., & Jang, H. –Y. (2005). Molecular Learning of wDNF Formulae, Preliminary Proceedings of the Eleventh International Meeting on DNA Computing (DNA 11). (to appear)

부 록

부록 1. 모델에서 사용된 기본 정서 표현 단어 목록

< symbol id: name: emotion value >

152: fun: happy
167: enjoy: happy
281: happy: happy
319: hate: disgust
575: afraid: fear
586: sad: sad
767: funny: happy
922: angry: angry
1097: fear: fear
1108: surprise: surprise
1546: anger: angry
1588: unhappy: disgust
1634: happiness: happy
1665: boring: disgust
1772: pleasure: happy
1765: boredom: disgust
1941: joy: happy
2093: excite: happy, surprise
2231: nervous: fear
2354: enjoyable: happy
2389: pleasant: happy
2574: wonder: surprise

2768: annoy: angry

2980: unpleasant: sad

3002: annoying: angry

3094: frighten: fear, surprise

3176: dislike: disgust

3326: rage: angry

3490: pleasing: happy

3571: embarrass: surprise

3657: delight: happy

3659: exciting: happy, surprise

3822: scary: fear

4182: sadness: sad

4462: undesirable: disgust

4700: amaze: surprise

4924: grief: sad

4933: disappoint: sad, angry

ABSTRACT

Affective aspects of communication are now recognized to be a crucial part of human intelligence, and together with several other emotional skills, have been argued to be more important for success in life. But human-computer interaction was far from emulating natural human-human interaction. This study suggested a computational model of recognizing emotion in text in this context.

This model generated semantic network based on the study that meaning of word and emotional recognition of word are closely related and inferred emotion with it. To simulate human cognition molecular computer is suitable in that its computation is massively parallel and this resembles biological signal processing system. Computation process with molecular computer is self-assembly process that DNA molecules make bonds to generate double-stranded molecules. Computation result is calculated statistically from distribution of self-assembled molecules by means of PLM algorithm. Chemical processing of word-molecules and linker-molecules is like to simple deductive reasoning, in a broader view it is probabilistic evaluation process with a semantic network.

Texts are preprocessed and converted to word bag. Model takes this word bag as input and calculates emotion value from relevant words retrieved by semantic network.

Keyword: molecular computer, PLM, recognizing emotion, semantic network

감사의 글

작고 보잘것없는 논문이지만 항상 곁에서 지켜봐 주시며 분에 넘치는 사랑을 주셨던 고마우신 분들께 감사의 마음을 전하고자 합니다.

먼저 논문을 지도해주신 장병탁 선생님께 감사드립니다. 2년의 짧은 시간이었지만 선생님의 자상하신 가르침은 늘 부족했던 저에게 용기를 불어넣어 주었습니다. 학자로서 선생님께서 보여주셨던 끝없는 열정과 성실함은 마음속 깊이 새겨서 영원한 마음의 등불로 삼겠습니다. 심사위원을 맡아주신 고성룡 선생님과 신효필 선생님께도 감사드립니다. 바쁘신 와중에도 제 학위논문 심사를 흔쾌히 맡아주셨을 뿐만 아니라, 제 연구에 많은 관심을 가져주시고 아낌없는 조언을 해주셨습니다.

인지과학 협동과정의 동료들과 선배님들, 그리고 후배들에게 감사드립니다. 문정 누나, 종호 형, 윤현 형, 소영 누나, 성호 형, 은석 형, 수동 형, 영균 형, 현애, 산, 지수, 지현, 윤정, 은정, 아림, 미선, 형주. 함께했던 소중한 시간들은 제게는 무엇과도 바꿀 수 없는 값진 시간들이었습니다. 항상 든든한 버팀목이 되어주신 여러분들께 다시 한번 감사의 마음을 드립니다.

오랜 친구들에게도 감사의 마음을 전합니다. 동환, 민택, 성찬, 성환, 원곤, 창규. 이들의 따뜻한 관심과 끊임 없는 애정이 지치고 힘들 때마다 저에게 큰 힘이 되어 주었습니다.

마지막으로 묵묵히 뒤에서 절 지켜봐 주시며 아낌없는 성원을 보내주신 부모님께 감사를 드립니다. 그 무엇으로도 표현할 수 없는 부모님의 관심과 사랑이 없었다면 지금의 저는 있을 수 없었을 것입니다. 앞으로도 두 분의 사랑에 감사하며 열심히 살아가겠습니다. 이 논문을 두 분께 바칩니다.